MJEN

MANAS Journal of Engineering
ISSN 1694-7398 | e-ISSN 1694-7398

Volume 9, Issue 1, (2021) Pages 45-54
https://doi.org/10.51354/mjen.869736

One Step Ahead Prediction of Ozone Concentration for Determination of Outdoor Air Quality Level

Waleed Khalid M. Mahmood[1,*], Ercan Avşar[2]

[1]Çukurova University, Department of Electrical and Electronic Engineering, Adana, Turkey, waleed.mahmood93720@gmail.com, ORCID:0000-0002-4973-0106
[2]Dokuz Eylül University, Department of Computer Engineering, İzmir, Turkey, ercan.avsar@deu.edu.tr, ORCID: 0000-0002-1356-2753

ABSTRACT

With the rapid spread of urbanization, competent authorities become increasingly anxious from air pollution risks and effect on citizens especially those with respiratory diseases. In this work, performances of six machine learning methods were analyzed for prediction of maximum ozone (O_3) concentration for the next-day. The models make the prediction using concentrations of six atmospheric components (PM2.5, PM10, Ozone (O3), Sulfur Dioxide (SO2), Nitrogen Dioxide (NO2), and Carbon Monoxide (CO)). The utilized machine learning methods are multilayer perception (MLP), Support Vector Regression (SVM), k-Nearest Neighbor (K-NN), Random Forests (RF), Gradient Boosting (GB), and Elastic Net (EN). After the predictions made by these models, the predicted values were further processed to be classified into one of the six air quality levels defined by United States Environmental Protection Agency. The prediction performances of the models as well as their corresponding classification results were analyzed. It was shown that MLP model gives the lowest RMSE of 2246 for prediction step while SVR achieved the highest accuracy score of 0.790.

ARTICLE INFO

Research article

Received: 28.01.2021
Accepted: 16.04.2021

Keywords:
machine learning,
time-series forecasting,
regression methods,
sequence-to-sequence,
air quality index.

[*]Corresponding author

1 Introduction

The increase in industrialization and urbanization has effects on the natural balance, this consequently affect human, animal, and plant wellbeing. Air pollution is one of the results of this situation and is considered to be the fourth of the biggest killer in the world after tobacco, high blood pressure and poor diet [1]. The surveys confirm the possibility of an increased number of deaths 40% from outdoor air pollution than today, therefore the United Nations Sustainable Development to aim to put strategy for the next two decades to limit air pollution hazards [2]. The Organization for Economic Cooperation and Development (OECD) has expected in 2012 that Outdoor Air Pollution will endure increasing towards 2050, and eventually will be the main cause of environmentally-related deaths worldwide [3].

Ambient Air Pollution is one of the mainsprings that worsens respiratory disease like Asthma and Chronic Obstructive Pulmonary Disease (COPD); therefore, air pollution has classified as a threat factor for people with these disease. According to the Global Asthma Network (GAN) studies, there are approximately 334 million asthma patients worldwide and accounts for 383 thousand deaths annually [4]. The World Health Organization (WHO) in 2016 has published an estimate survey that ambient air pollution accounts for 4.2 million premature death worldwide and more than 18% of deaths were courses COPD [5, 6].

According to the surveys, outdoor air pollution considers the major of many substantial adverse to universal public health, therefore, the level of hazard requires the cooperation of all researchers to study, analyze and reduce the risk of outdoor air pollution on the health of chronic respiratory patients [7]. In addition to civilized awareness, people are becoming more attentive due to changes in air quality and their impact on respiratory patients health. Although the children and young adults spend most of the time indoors because of COVID-19, the air quality-related with outdoor environments is still important for Asthma and COPD patients because outdoor environment air has a major impact on the indoor air quality. The reason for such an impact is that there is a lack of efficacy in preventing particulate matter transmission from the outside environment to the inside environment [8].

The significance of air quality forecasting is increasing rapidly for establishing citizen’s superior quality and safety life to adapt with growing and developing of the modern era, and for providing an influential and appropriate environment for sensing research. Therefore, this work aims to explain the potential ways to set up a time series forecast model architecture, which to facilitate the establishment of an airtight future model aim to support respiratory disease patients.

Majority of the previous studies rely on six important pollutants of the atmospheric components (PM2.5, PM10, Ozone (O3), Sulfur Dioxide (SO2), Nitrogen Dioxide (NO2), and Carbon Monoxide (CO)) to health problems. O3 is formed by complex chemical processes having a high impact on the environment. As a result of O3 exposure, approximately more than 21,000 premature deaths were reported in Europe, and 1.1 million in worldwide premature deaths, many of death have respiratory diseases [9]. Therefore, O3 is selected as the target pollutant to be predicted in this study. For this purpose, six regression models that are artificial neural networks (ANN), support vector machines (SVM), k-nearest neighbor (KNN), random forests (RF), gradient boosted decision tree (GBDT), and elastic net (EN), utilized for establishing a time series-forecasting model. The prediction models were trained by using several measurements of the pollutant gases and the daily maximum O3 level for the next day was forecast using these models. For the classification method, the predicted values obtained by training data were classified to five levels of air quality (good, moderate, unhealthy for sensitive groups, unhealthy, very unhealthy). United States Environmental Protection Agency (US EPA) standard equation was used calculate the air quality level and hence to classify the predicted value. These classification results were used to generate a confusion matrix, eventually.

2 Literature Review

In recent times, air pollution threats have been studied widely by the researchers, several of them have contributed to investigate interesting solutions with the potential of reducing the risks. Most of them consider machine-learning methods as the ideal solution to develop a time series forecast model. In this section, we will discuss some of the machine learning methods of previous work.

In earlier years, many researchers have tended to compare regression models. For example, N. K. Ahmed et al. presented a paper to compare the major regression models for building the architecture of time-series forecast [10]. The comparisons of models are substantial movement to highlight the appropriate model to establish a time-series forecast, also, provide solutions to time-series forecasting problems. In 2010, Ping-Feng Pai et al. illustrated the significance of using SVM with time series forecasting and compared the performance with Autoregressive Integrated Moving Average (ARIMA) [11]. Involvement of SVM in building the architecture of sequence-to-sequence forecast provide a major effect on the generalization performance, also doesn't take much time to implement. The substantial common solution in the growth era is establishing a hybrid model to eliminate the main problems in time-series forecast. In 2013, F. S. de Albuquerque Filho et al. suggested an intelligent hybrid system for time-series forecasting to predict the levels of pollution of four components (CO, SO2, PM10, and NO2) in the atmosphere. The hybrid model built by an ANN-MLP with a particle swarm optimization (PSO) algorithm [12]. ANN-MLP model considers an appropriate model for time-series forecast, however, it is a good idea to reduce errors in a model by using a provide potential in PSO algorithm and combined with MLP. Concerning regression tree methods, James R. Lloyd clarified a susceptibility to perform time-series methods by using a GD model to forecast hourly loads of a US utility [13]. Then, H. Tyralis and G. Papacharalampous focused on clearly the potential of using an RF model on time-seres forecasting [14]. Regression tree methods have proven competent to time-series forecasting, however, in future are expected to provide better prediction accuracy.

Due to rapid of industrialization and economic boom in the urban cities of worldwide, Sarajevo capital city of Bosnia has suffered from a high level of pollution during the winter session. In 2016, Maja M. Dedovic et al. presented a research paper urges to reduce the accumulated risks to Sarajevo by recourse to sequence-to-sequence forecast method. They suggested a model predict the concentration of pm10 particles from 2010 to 2013 together with the environmental measures like temperature, humidity, wind speed and pressure. The proposed model depended on an ANN model for training a dataset. In model evaluation, the proposed model output a forecasting result with R^2 from 0.5 to 0.9 of total years [15]. Bing-Chun Liu et al, proposed an AQI forecast model to predict air pollution in three urban cities in China. The main goal of the proposed model is an enhancement of air quality forecasting results via minimizing the errors of machine learning prediction algorithms. The machine learning regression method used in the proposed model is SVM for predict air quality index. They illustrated suffers, problems and struggles with air pollution for each city. The proposed regression model achieved high-performance efficiency, besides, MSE results achieved by training model is between 106.22 to 128.70 of total cities [16]. In 2018, Hong Zheng et al. presented a paper highlighting on rising of air pollution and reduce its impact in China. proposed a new approach of multiple kernel learning (MKL) with support vector classifier (SVC) to train air quality dataset. MKL-SVC performance compared with various machine learning models had widely used, like SVM, RF, ANN-MLP, ARIMA, and long short-term memory (LSTM). The proposed model shown high performance beyond the other models, also they praised SVM, RF, and ANN-MLP performance these presented very good performance than the sequence-to-sequence models LSTM and ARIMA. MKL-SVC model succeeded accuracy of 0.972 with MSE of 0.030 [17]. F. Martinez et al. illustrated the potential to train a time-series dataset by KNN model. It is possible to use KNN model in time series architecture, however, it is not the most suitable among the previously implemented models [18]. In 2020, K.Maheshwari and S. Lamba, they proposed time series forecast model. The proposed model aimed to minimize the effects of air pollution emission. the proposed model forecast future concentrations of pm2.5 depend on six machine learning regressors models for training a dataset, were Linear Regression, ANN-MLP, K-NN, Decision Tree, RF and Stochastic Gradient Descent. The proposed model achieved an accuracy of 95.5% [19]. F. Shen et al. proposed a prediction model to overcome the limitation in time-series forecasting. The combination of EN model and high order Fuzzy cognitive maps (FCMs) shown predictable of a time-series dataset with less error compared with other regression models [20]. EN model is a combination of lasso regression and ridge regression, it has proven efficiency of predict time series compare with lasso and Ridge.

3 Materials and Methods

3.1 *Time Series Dataset*

Dataset was obtained from The World Air Quality Index project; they have published it with the spreading out of COVID-19 [21]. Dataset Includes 380 major cities in the worldwide, has been collected from the several stations since 2016 until now it is updated daily. The World Air Quality Dataset contains the minimum, the maximum, the median and the standard deviation for daily measurements of the six air pollutant species ($PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3). Among these species, the unit for O_3 and CO is parts per million (ppm), while SO_2 and NO_2 are measured in parts per billion (ppb). On the other hand, the unit for the particulate matter pollutants, PM_{10} and $PM_{2.5}$ is micrograms per m^3 ($\mu g/m^3$). The atmospheric components like temperature, humidity, wind speed, wind gust, dew, and pressure are not included in the dataset, hence not utilized in this study. A sample part of the dataset is provided in Table 1. In this work, Istanbul was chosen as the city whose air quality to be predicted for the next day. After pre-processing steps, the dataset size has obtained consists of 1013 row and 56 columns. The data between the dates from 26.12.2016 to 14.10.2020 is used as the dataset. The reason for selecting this time interval is to be able to include the data belonging to covid-19 situation. In that case, the future expansions to the current study may cover detailed analysis about the temporal effects of covid-19 on the outdoor air pollution.

Table 1. An illustrative table of dataset contents.

Date: From 12/26/2016 To 10/15/2020	**Country: Turkey**	**City**	**Species**	**Air Pollutant Species (S)**			
				Min	**Max**	**Median**	**Variance**
12/26/2016 12:00:00 AM	TR	IST	SO_2 (ppb)	0.6	20.8	3.6	150.56
12/26/2016 12:00:00 AM	TR	IST	PM_{10} (ppm)	1	119	32	2965.21
12/26/2016 12:00:00 AM	TR	IST	O_3 ($\mu g/m^3$)	0.5	18.3	2.5	136.67
12/26/2016 12:00:00 AM	TR	IST	NO_2 (ppb)	6.9	76.1	28.8	2093
12/26/2016 12:00:00 AM	TR	IST	CO ($\mu g/m^3$)	0.1	59.3	7.6	1997.7
12/26/2016 12:00:00 AM	TR	IST	$PM_{2.5}$ (ppm)	2	751	85	61407.9

3.2 *Air Quality Index*

After performing prediction by the models, the predicted values are assigned an AQI level and then compared with the AQI level of the actual value. Air Quality Index (AQI) is a required index that represents air quality based on air pollutant concentrations at a certain period. AQI schedule varies depending on the country, each country containing a specific Index. AQI Accredited by United States Environmental Protection Agency (US EPA) standard has divided into 6 groups. Each group symbolized by different colors and a standardized public health advisory depending on the concentration of air pollution [22]. The US EPA standard method for calculating AQI is given in Equation (1) and the levels for AQI indexes are illustrated in Table 2. The green color indicates the best and changes gradually to maroon indicating the worst. The orange color has been reported to be the threshold for risky situations for Sensitive Groups (lung disease, older adults and children).

Equation (1): The US EPA standard equation to calculate air quality index.

$$I_p = \frac{I_{HI}-I_{LO}}{BH_{HI}-BH_{LO}}\left(C_p - BH_{LO}\right) + I_{LO} \qquad (1)$$

Where I_p= the index for pollutant p

C_p= the truncated concentration of pollutant p

BH_{HI} = the concentration breakpoint that is greater than or equal to C_p

BH_{LO}= the concentration breakpoint that is less than or equal to C_p

I_{HI}= the AQI value corresponding to BH_{HI}

I_{LO}= the AQI value corresponding to BH_{LO}

Table 2. AQI levels.

AQI	Air Pollution Levels	Health Implications
0-50	Good	Air quality is satisfactory, no risk.
50-100	Moderate	Air quality is acceptable.
100-150	Unhealthy for Sensitivity Group	Members of sensitive groups may experience health effects.
150-200	Unhealthy	Members of the public may experience health effects, for sensitive groups will be more seriouse health effects.
200-300	Very Unhealthy	The risk of health effects is increased for everyone; health alert.
300-500	Hazardous	Health warning for emergency conditions.

3.3 Time-Series Dataset Preprocessing

Initial step, Istanbul city dataset was extracted from the public main dataset, and then the six essential components for atmospheric pollution are separated from other atmospheric components. Eventually, a time-series data containing minimum, maximum, median, and variance of daily measurements of six pollutants for 37 months of duration is obtained. This dataset was prepared for the time-series prediction task by performing the steps explained in the following subsection.

3.3.1 One-Step Ahead Forecast

The purpose of time-series prediction is estimating the future information by past and current information samples. Therefore, one-step forecast strategy (in this problem it may be defined as "next day predict strategy") is a supervised learning task. The main goal of this strategy is generating forecasting dataset from current data by reframing the original dataset to derive historical dataset (t-1, t-2,.….t-n) and next-day (t+1) dataset, then merge all of them to gain entry time-series dataset with target dataset, therefore, predicting the future values (t+1, t+3,…. t+ n) (Figure 1).

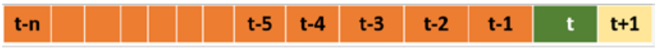

Figure 1. *One-step Ahead strategy.*

3.3.2 One-hot Encoding

The main step in preprocessing dataset is a coding method, for getting the better results in machine learning (ML) based forecasting. One-Hot Encoding was used to encode time-series dataset. One-Hot Encoding employs N-bits status, each independent category of entry has one bit. Bits represented by "one" or "zero" and only one of the entries will be valid with a state of "one". As shown in Table 3, One-Hot Encoding was used to encode weekdays, a 7-bits encoding had used to differentiate each day. For example, for a sample that corresponds to measurements for Wednesday, the one-hot feature vector is 0001000.

Table 3. *One-Hot Encoding scheme for weekdays.*

days	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
12/25/2016	**1**	**0**	**0**	**0**	**0**	**0**	**0**
12/26/2016	**0**	**1**	**0**	**0**	**0**	**0**	**0**
12/27/2016	**0**	**0**	**1**	**0**	**0**	**0**	**0**
12/28/2016	**0**	**0**	**0**	**1**	**0**	**0**	**0**
12/29/2016	**0**	**0**	**0**	**0**	**1**	**0**	**0**
12/30/2016	**0**	**0**	**0**	**0**	**0**	**1**	**0**
12/31/2016	**0**	**0**	**0**	**0**	**0**	**0**	**1**

3.3.3 Standardization

In general, the machine learning models require to rescale input dataset to be sure that the input features are measured on the same scale, especially when all data on a completely different scale. Standardization or called Z-score normalization refers to transform input dataset to zero mean and unit standard deviation to be on the form of a normal distribution. The formula for applying z-score normalization is given in Equation (2) where X denotes the input feature, µ and $\boldsymbol{\sigma}$ are mean and standard deviation of the feature, respectively. This transformation is applied to all features other than one-hot encoded binary features.

$$\boldsymbol{z} = \frac{x - \mu}{\sigma} \tag{2}$$

3.3.4 Walk-Forward Validation

The common methods in machine learning such train-test split and k-fold cross-validation are avoided here because it is not a valid approach for time series datasets. The dataset was divided into training and testing subsets based on train test split, which esteems the temporal order of observations. To eliminate common errors in time series forecasting, the prediction of test set should not rely solely on the training set. Common errors were avoided by performing Walk-Forward validation methods. Walk-Forward validation approach is predicting one value from the test set for each iteration at a time. After predicting the value taken from the test values will add this value to training set or widow set (number of training data each iteration) for the next iteration. Training and predicting of data that will continue until the end of all iteration steps, consequently, the number of iteration should be equal to the number of tests. There are two types of Walk-Forward validation, Rolling Walk Forward and Anchored Walk Forward. In this study, Rolling Walk Forward (also named as rolling window) method is used. Rolling window approach is the in-sample periods "roll"; meaning that the window size is steady during all iteration steps [23]. In other words, the number of samples in the training set will not change during the prediction process, as shown in Figure 4.

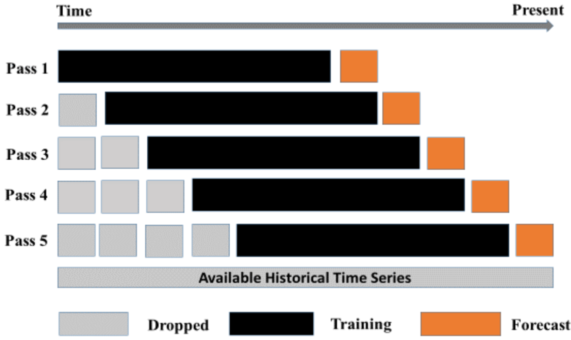

Figure 2. *Illustration Rolling Walk Forward validation.*

3.4 *Regression Models*

Six regression models were used to perform the time series forecasting. The forecasting models used a past and current time step to next-day forecast, in turn, to compare with the future data we have already specified as preliminarily target sets. The comparison for measurement of our proposed model's performance with the MSE and R² for regression methods, and accuracy, F-score, Recall, and Precision for classification methods. In this section, we will clarify applied regression models for the sequence-to-sequence forecast**.**

3.4.1 Support Vector Regression

Support Vector Machines (SVMs) are a satisfactory way to perform on regression problems based on Vapnik-Chervonenkis (VC) theory [24], it evolved over the last decade to generalize to become appealable on yet-to-be-seen data. As a supervised machine learning methods, it has been proven to be most effective in a real time-series forecast [25]. The prediction functions $f(x)$ for linear and non-linear regression applications had derived from one-dimensional example shown in equations (3, 4). If we assume that $f(x)$ is non-linear, thus, our aim is mapping data to higher dimensional "feature" space or called kernel space, which achieved by Kernel function $\phi(x)$ [26]. Kernel functions types are Gaussian function and polynomial, and hyperbolic tangent. For our proposed model chosen Gaussian function or radial basis function (RBF), as shown in Equation (5), where$\|\boldsymbol{x} - \mu\|$ is the squared Euclidean distance of two-feature vector and $\boldsymbol{\gamma}$ is gamma. We aim to find the best weight (w) and threshold (b). Parameter C is the regularization parameter setting the margin of the decision function in linear and non-linear SVM. If we encourage a large margin, should select a small value of parameter C, consequently will get a lower misclassification rate and vice versa.

$$\boldsymbol{f(x) = (w.x) + b} \tag{3}$$

$$\boldsymbol{f(x) = (w.\phi(x)) + b} \tag{4}$$

$$\boldsymbol{\phi(x, \mu) = exp(-\gamma\|x - \mu\|^2)} \tag{5}$$

3.4.2 K-Nearest Neighbors

K-Nearest Neighbors (KNN) regression is one of lazy learning or instance-based learning (memory-based learning) that use for time-series forecast. KNN is a simple non-parametric algorithm that based on determining the most k- nearest or similar of new samples (test instance samples with an unidentified target) to the training instance (feature and its target samples) according to the similarity measure or a distance metric [27]. The k parameter represents the number of neighbors to consider for determining the predicted value of a test sample. For KNN regression, the average of closest k samples is calculated as the prediction. Equation (6) shows the mathematical equation that used to calculate the distance metric.

$$\mathrm{d}_{\mathrm{ij}} = \sqrt{\textstyle\sum_{m=1}^{n}\left(\mathrm{x}_{\mathrm{im}} - \mathrm{x}_{\mathrm{jm}}\right)^2} \tag{6}$$

Where d_{ij} represent the distance metric between specified dataset, m is the specific time, x_{im} is the training instances, i, j is the dimensionality of the feature space and x_{jm} is the testing data with an unidentified target.

3.4.3 Multilayer Perceptron

Fundamentally, Artificial Neural network (ANN) is a group of nodes or units connected with each other like neurons in a human brain, each node is able to transfer data to the other starting from the input layer going through the hidden layer, and reaching the output layer. Multilayer Perceptron (MLP) is an ANN belong to feedforward neural network class, which means the input data move to only forward directions [28]. MLP has a wide agreement that it is the appropriate model for time-series forecasting regression because it provides a non-linear tool for resolving regression problems, and is able to approximate any smooth function without prior presumption about data distribution. MLP model was tuned by adjusting number of hidden layer units, which is the first and most effective parameter that permits us to set a number of layers and nodes possessed by the neural network.

3.4.4 Regression Trees

Regression Trees (RT) are based on the principle of splitting the input parameters space into special independent and non-overlapping regions according to set rules. Many alternatives work on similar prediction principle with better predictive performance and a completely different way to create trees are random forest (RF), and gradient boosting (GB). RF algorithm approach is forming trees independently, with maximum depth. RF structure permits to minimize the variance of a huge number of complex decision trees (high depth) [29]. GB is a numerical optimization algorithm that works on approach adding a new decision tree at each step iteratively to minimize the loss function [30]. The RF and GB regression models was tuned by adjusting the number of trees, which is the major effective parameter that determine the number of sequential trees for best forecasting.

3.4.5 Elastic Net

Elastic net is a linear regression model that considers a combination of lasso regression (least absolute shrinkage and selection operator) and Ridge regression. Lasso is a penalized minimal squares method proposed by Tibshirani (1996), which imposes an L1-penalty on the regression coefficients [31]. The aim of EN method is reducing the loss function that given below:

$$L_{en}(\widehat{\boldsymbol{\beta}}) = \frac{\sum_{i=1}^{n}\left(y_i - x_i^J\widehat{\beta}\right)^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j-1}^{m}\widehat{\boldsymbol{\beta}}_j^2 + \alpha\sum_{j-1}^{m}\left|\widehat{\boldsymbol{\beta}}_j\right|\right) \tag{7}$$

Where λ is the model parameter, **x** is the input matrix, **y** is the target vector and $\boldsymbol{\beta}$ is the weight vector. Alpha (α) is the mixing parameter that combines Ridge and Lasso methods. When $\alpha = 0$ the elastic net will corresponds to simple ridge regression, while $\alpha = 1$ to lasso regression.

3.5 *Performance metrics*

In order to predict the air quality level for the next day, fist the ozone concentration is first predicted then this predicted value is assigned a level using Equation 1. Hence, different performance measures are calculated for each of these steps. Since the first step is a regression problem, root mean squared error (RMSE) and R^2 values are calculated according to the equations given below:

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_{Ti} - Y_{pi}\right)^2}{n}} \tag{8}$$

Where n is number of non-missing data point, Y_{Ti} and Y_{pi} denote the observed and predicted values, respectively.

$$\boldsymbol{R^2} = 1 - \frac{RSS}{TSS} \tag{9}$$

Where R^2 is the coefficient of determination.

RSS is the residual sum of squares = $\sum_i\left(Y_{Ti} - Y_{pi}\right)^2$ (10)

TSS is the total sum of square = $\sum_i(Y_{Ti} - Y_{mi})^2$ (11)

And Y_{mi} is denote the number of the Y_{Ti}.

The assigned level of air quality is then compared by its actual level. In order to do it for all the predictions, a multi class confusion matrix is generated because five different levels of air quality are defined. A generic confusion matrix considering k classes is given in Figure 3.

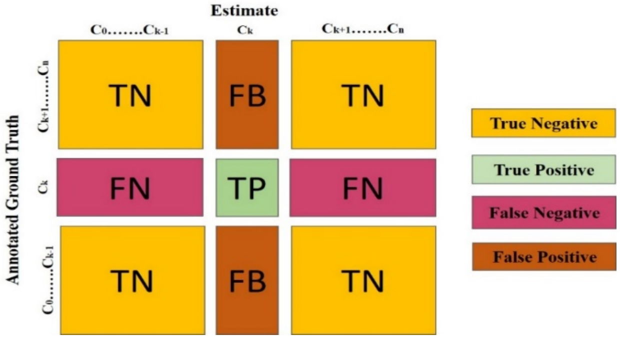

Figure 3. *The confusion matrix for multi-class.*

For the final classification part, accuracy, precision, recall, and f1-score are calculated according to the equations given below:

$$\boldsymbol{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \tag{12}$$

$$\boldsymbol{Precision} \;=\frac{TP}{TP+FP} \tag{13}$$

$$\boldsymbol{Recall} \;=\frac{TP}{TP+FN} \tag{14}$$

$$\boldsymbol{F1\ Score} \;=\frac{2*(Recall*Precision)}{(Recall+Precision)} \tag{15}$$

4 Results and Discussion

In the previous section, the dataset size and concerned region were mentioned, and the methods were implemented to process the dataset as well as the architecture of the proposed algorithm for the time-series forecasting. In this section, the results obtained with the mentioned methods are presented. Since these methods have different parameters that influence the prediction performance, a gird of parameters for the methods are used in testing of the methods. These parameters are given in Table 4. For the number of neighbors (k) in KNN model, three values are used and the lowest RMSE was obtained when K=3. In the SVM model stage, the linear method is shown with regularization parameter equal one achieved the best RMSE result compared with the non-linear method. As for MLP, several of hidden layer standard and the maximum number of iteration was examined. In an MLP with two hidden layers, where the first layer has 14 units and the second has 7 units, yielded the smallest RMSE when the maximum training iterations was set to 300. In the EN model, three alpha values were examined and 0.9 was found to be the most suitable one compared with others. As for the number of trees in regression trees models, three values are examined for both RF and GB, 100 and 300 are determined to be the best number of trees for RF and GD, respectively. The obtained RMSE and R^2 for the regression methods with the best parameters are given in Table 5. As indicated earlier, the predicted values are categorized into different levels of air quality and then compared with the actual level. The results associated with this classification step is given in Table 6.

Table 4. *The examined methods and the corresponding parameters.*

Method	Parameters	Tried values for the parameters
KNN	K	3,5,7
SVM	Kernel type	Linear, Gaussian (RBF)
	C	1,3,7,550,700
	gamma	0.001, 0.0008, 0.0003
RF	Number of estimators	100, 300, 500
EN	Alpha	0.1, 0.5, 0.9
MLP	Number of maximum iterations	200, 300 , 400, 500 , 700, 1000
	Hidden layer size	(10,5), (14,7), (20,10), (50,25), (25,), (50,), (70,), (100,), (150,), (200,), (250,), (300,)
GB	Number of estimators	300, 500, 700

Table 5. *Regression results for air quality prediction.*

	KNN	SVM	RF	EN	MLP	GB
RMSE	2594	2284	2548	2250	**2246**	2745
$\boldsymbol{R^2}$	0.45	**0.52**	0.46	**0.52**	**0.52**	0.42

Table 6. *Classification results for air quality prediction.*

	KNN	SVR	RF	EN	MLP	GB
Accuracy	0.756	**0.790**	0.713	0.726	0.697	0.709
recall	0.756	**0.790**	0.713	0.726	0.697	0.709
F1-score	0.757	**0.788**	0.736	0.743	0.716	0.732
Precision	0.76	**0.786**	0.772	**0.786**	0.765	0.760

The performance measures for regression models in Table 5 are close to each other. However, it is clearly seen that MLP performs the prediction with lowest error and three methods (SVM, EN and MLP) has the highest R^2 score. In Figure 4, the actual and predicted values on the test set are given on the same plot for visual comparison. In addition, these values are presented as a scatter plot in Figure 5. Both of these figures depict the results obtained by MLP model, which outputs the lowest RMSE value. It is obvious from these figures that the majority of errors are caused by the samples where the actual O_3 concentration is very high. This is probably due to the fact that the samples with high O_3 value are very rare. As a result, it becomes difficult for the learning methods to model the samples with such behavior. On the other hand, the classification results in Table 6 show that the accuracy and recall values for all of the methods are identical. This implies balanced classification models, meaning that the ability of the models for classifying the positive samples and the negative samples are the same. This may be due to the indirect class assignment of the test samples. In other words, since the corresponding O_3 values are predicted for test samples firstly and next, the corresponding class labels are calculated according to Equation 1, the classification results may have such a balance. Besides, it notable that predictions obtained by SVR method yield the highest classification result for all of the metrics. The corresponding confusion matrix for SVR predictions is given in Figure 6.

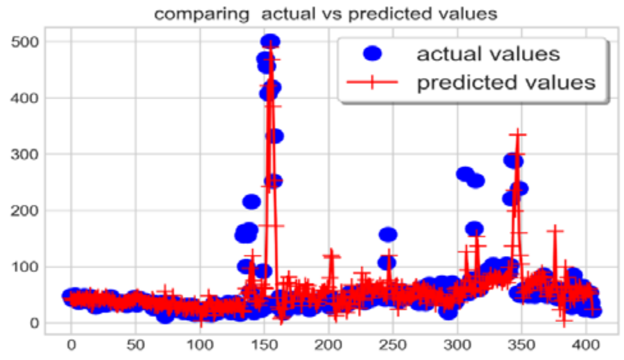

Figure 4. *Actual and predicted values for MLP model*

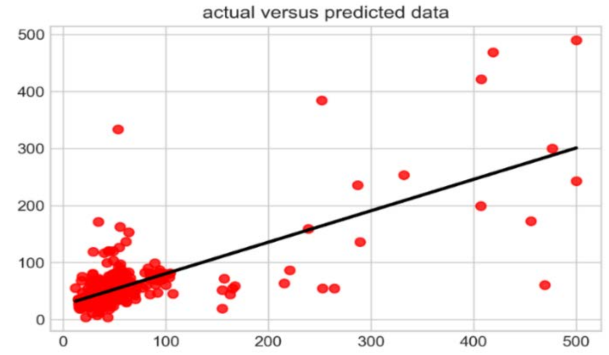

Figure 5. *Actual versus predicted values for MLP models in a scatter plot*

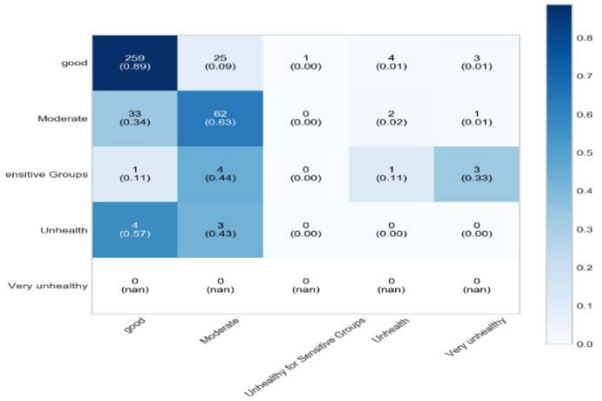

Figure 6. *The confusion matrix for SVR predictions. Rows and columns indicate actual and predicted values, respectively.*

In the related literature, there are other studies in which the future values of air pollutant concentrations are predicted. Majority of these studies try to solve this problem by means of regression methods only. In other words, no further calculations for air quality level assignment are made in these works. Furthermore, only a limited number of methods with predefined parameters are used in these works. For example, P. García Nieto et al. presented a prediction model for forecasting the time-series dataset of PM based on SVM, VARMA, ARIMA and ANN-MLP [32]. The lowest RMSE measure was achieved by SVM as 2.061. In the study proposed by Z. Meng, a prediction model to forecast ground O_3 level based on machine learning methods [33]. SVM, decision tree, RF, and logistic regression models were utilized to train the O_3 dataset which contains binary target values as high or low ozone concentration. In other words, no regression model was fit on the data and the obtained classification accuracies ranged between 0.8 and 0.949. Another study about classification of air quality levels considers five different stages for the AQI [34]. The study uses Naïve Bayes and decision trees for the classification task where 91.99% of accuracy was observed with decision trees. Even though the accuracy score is higher than those obtained in this work, the F1-score performance was not reported. Therefore, it becomes impossible to analyze the performance in case of a class imbalance problem in the dataset. In a recent study, ANN method is utilized for predicting four pollutant species that are NO_2, PM_{10}, $PM_{2.5}$, and O_3. The results were enhanced through an operation called real-time-corrections and it was shown that prediction performance of O_3 can have RMSE and R^2 scores of 0.187 and 0.79, respectively [35]. However, in that work, no information regarding the unit of the O_3 measurement is provided. Therefore, the RMSE results is not comparable with those obtained in this study. In addition, performances of different machine learning methods have not been compared and suitable parameters for the ANN model has not been searched in that work. Even though there are some studies in which performances of various methods have been analyzed and compared together [36, 37], a two-step method (class assignment followed by regression) for air quality level detection has not been proposed earlier.

5 Conclusion

In this paper, several machine learning methods were analyzed to predict maximum O_3 concentration for the next-day as an indicator of outdoor air quality. The predictions were made for one-step ahead in the test set and the training was performed using rolling walk forward validation method. Therefore, a distinct training process was performed for each test sample. Total of six different machine learning methods, namely, KNN, SVM, RF, EN, MLP, and GB were used for time-series prediction experiments. Next, the predicted values were assigned a level for the air quality. This assignment was made using the formula proposed by United States Environmental Protection Agency. As a result, the problem was transformed into a multi-class classification problem. For the first prediction step, MLP was observed to achieve the lowest error while SVR predictions were better classified into the air quality levels. Therefore, it may be concluded that low prediction accuracy does not always imply better representation for air quality levels. Another noticeable point in the predictions is that error related with low pollution values is lower but the models make more error for the samples corresponding to high pollution values. This means that increasing the number of samples for high pollution measurements in the dataset may enable the models learn the patterns for such samples.

References

[1] Niall McCarthy, “Air Pollution Contributed to More Than 6 Million Deaths in 2016, Data journalist covering technological”, societal and media topics, 2016.

[2] P.Rafaj, G.Kiesewetter , T.Gul, W.Schoppa, J.Cofala, Z.Klimont, P.Purohit, C.Heyes, M.Amann, J.Borken-Kleefeld, L.Cozzi. Outlook for clean air in the context of sustainable development goals. : Global Environmental Change, September 2018.

[3] OECD (2012), OECD Environmental Outlook to 2050, OECD PUBLISHING. https://dx.doi.org/10.1787/9789264122246-en.

[4] Aloke Ghoshal, Pradyut Waghray, George Dsouza, Mahip Saluja, Mayank Agarwal, Ashish Goyal, Sneha Limaye, Akash Balki, Sudhir Bhatnagar, Manish Jain, Sharad Tikkiwal, Abhijit Vaidya, Meena Lopez, Rashmi Hegde, Jaideep Gogtay, “Real-world evaluation of the clinical safety and efficacy of fluticasone/formoterol FDC via the Revolizer in patients with persistent asthma in India”, On 25 November 2019, 10.1016/j.pupt.2019.101869.

[5] Burden of disease from ambient air pollution for 2016, 1211 Geneva 27, World Health Organization 2018. https://www.who.int.

[6] World Health Organization, Ambient air pollution: a global assessment of exposure and burden of disease, 2016, https://apps.who.int/iris/handle/10665/250141.

[7] X. Li, Ling Jin, and H. Kan, Air pollution: a global problem needs local fixes, 25 JUNE 2019, china, https://doi.org/10.1038/d41586-019-01960-7.

[8] Blondeau, P., Iordache, V., Poupard, O., Genin, D., Allard, F., 2005. Relationship between outdoor and indoor air quality in eight French schools. Indoor Air 15, 2–12, 10.1111/j.1600-0668.2004.00263.

[9] Brian S. Freeman, G. Taylor, B. Gharabaghi, and Jesse Thé, forecasting air quality time series using deep learning, 24 May 2018. https://doi.org/10.1080/10962247.2018.1459956.

[10] Nesreen K. Ahmed, Amir F. Atiya , N.El Gayar &H. El-Shishiny, An Empirical Comparison of Machine Learning Models for Time Series Forecasting, 15 Sep 2010. https://doi.org/10.1080/07474938.2010.481556.

[11] Ping-Feng Pai, Kuo-Ping Lin, Chi-Shen Lin, and Ping-Teng Chang, Time series forecasting by a seasonal support vector regression model, June 2010, https://doi.org/10.1016/j.eswa.2009.11.076.

[12] Francisco S. de Albuquerque Filho, Francisco Madeiro e Sérgio M. M. Fernandes, Paulo S. G., de Mattos Neto, and Tiago A. E. Ferreira, Time-series forecasting of pollutant concentration levels using particle swarm optimization and artificial neural networks, Paulo 2013. http://dx.doi.org/10.1590/S0100-40422013000600007.

[13] James R. Lloyd, GEFCom2012 hierarchical load forecasting: Gradient boosting machines and Gaussian processes, 16 August 2013. https://doi.org/10.1016/j.ijforecast.2013.07.002.

[14] H.Tyralis, and G.Papacharalampous, Variable Selection in Time Series Forecasting Using Random Forests, 4 October 2017. https://doi.org/10.3390/a10040114.

[15] M. M. Dedovic, S. Avdakovic, I. Turkovic, N. Dautbasic, and T. Konjic, Forecasting PM10 concentrations using neural networks and system for improving air quality, 08 December 2016, 10.1109/BIHTEL.2016.7775721.

[16] Bing-Chun Liu, A. Binaykia, P. 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. July 14, 2017, https://doi.org/10.1371/journal.pone.0179763.

[17] H.Zheng ,H. Li, X. Lu, and T. Ruan, A Multiple Kernel Learning Approach for Air Quality Prediction, 12 Jun 2018, https://doi.org/10.1155/2018/3506394.

[18] F. Martínez, M. P. Frías, F. Charte and A. J. Rivera, Time Series Forecasting with KNN in R: the tsfknn Package, December 2019, ISSN 2073-4859.

[19] K.Maheshwari and S. Lamba, Air Quality Prediction using Supervised Regression Model, 03 February 2020, 10.1109/ICICT46931.2019.8977694.

[20] Fang Shen, Jing Liu, and Kai Wu, Multivariate Time Series Forecasting based on Elastic Net and High-Order Fuzzy Cognitive Maps: A Case Study on Human Action Prediction through EEG Signals, 29 May 2020. 10.1109/TFUZZ.2020.299851.

[21] World's Air Pollution: Real-time Air Quality Index, http://waqi.info/.

[22] Eusebio Jarauta-Bragulat, Carme Hervada-Sala, Juan Jose Egozcue, Air Quality Index Revisited from a Compositional Point of View, published online: 23 May 2015 © International Association for Mathematical Geosciences 2015 .

[23] A. Sanjivanrao More, D.Sunil Ranaware, B. D. Wamane, and G. S. Salunkhe, Enhancement in Financial Time Series Prediction with Feature Extraction in Text Mining Techniques, Nov 2019, 2395-0056, International Research Journal of Engineering and Technology (IRJET).

[24] V. N. Vapnik, An overview of statistical learning theory, Sept. 1999.10.1109/72.788640.

[25] N. I. Pankevych, R.Sankar, Time Series Prediction Using Support Vector Machines: A Survey, 24 April 2009. 10.1109/MCI.2009.932254.

[26] M. Awad, R.Khanna, Efficient Learning Machines (chapter: Support Vector Regression, Pages 67-80), 27 April 2015. https://doi.org/10.1007/978-1-4302-5990-9

[27] F.Martínez, M. P. Frías, M. D. Pérez, and A. J. Rivera, a methodology for applying k-nearest neighbor to time series forecasting, 21 NOV 2019. https://doi.org/10.1007/s10462-017-9593-z.

[28] J. Kukkonen, L. Partanen, A. Karppinen, J. Ruuskanen, H. Junninen, M. Kolehmainen, H. Niska, S. Dorling, T. Chatterton, R. Foxall, and G.Cawle, extensive evaluation of neural network models for the prediction of NO2 and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki, 22 Aug 2003. https://doi.org/10.1016/S1352-2310(03)00583-1.

[29] S. Touzani, J. Granderson, and S. Fernandes, Gradient boosting machine for modelling the energy consumption of commercial buildings, Nov 2017. https://doi.org/10.1016/j.enbuild.2017.11.039.

[30] Max Kuhn, Kjell Johnson, Applied Predictive Modeling, New York 2013, https://doi.org/10.1007/978-1-4614-6849-3.

[31] Hui Zou, and Trevor Hasti, Regularization and variable selection via the elastic net, 09 March 2005, https://doi.org/10.1111/j.1467-9868.2005.00503.x.

[32] P. García Nieto, F. Sánchez Lasheras, E. García-Gonzalo, F. de Cos Juez, PM10 concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: A case study, 2018, *Science of the Total Environment*, https://doi.org/10.1016/j.scitotenv.2017.11.291.

[33] Z. Meng, Ground Ozone Level Prediction Using Machine Learning, 2019, *Journal of Software Engineering and Applications*, 10.4236/jsea.2019.1210026.

[34] R. Waman Gore, D. S. Deshpande, An approach for classification of health risks based on air quality levels, 01 December 2017, India, 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), 10.1109/ICISIM.2017.8122148.

[35] S. Agarwal, S. R. Sharma, Md H. Suresh Rahman, et al., Air quality forecasting using artificial neural networks with real time dynamic error correction in highly polluted regions, 2020, Science of the Total Environment, https://doi.org/10.1016/j.scitotenv.2020.139454.

[36] Doreswamy, K. Harish Kumar, Y. Km, I. Gad, Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models, 2020, *Procedia Computer Science*, https://doi.org/10.1016/j.procs.2020.04.221.

[37] S. Ameer, M. Ali Shah, A. Khan; H. Song, C. Maple, S. Ul Islam, M. N. Asghar, Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities, 2019, 10.1109/ACCESS.2019.2925082.

[38] W. Leong, R. Kelani, Z. Ahmad, Prediction of air pollution index (API) using support vector machine (SVM), 2020, Journal of Environmental Chemical Engineering, https://doi.org/10.1016/j.procs.2020.04.221.

[39] Z. Yu, and Z. Niu; W. Tang, Deep Learning for Daily Peak Load Forecasting—A Novel Gated Recurrent Neural Network Combining Dynamic Time Warping, 29 January 2019, 10.1109/ACCESS.2019.2895604.

[40] C. S. Malley, D. K. Henze, Johan C.I. Kuylenstierna, H. W. Vallack, Y. Davila, S. C. Anenberg, M. C. Turner, and M. R. Ashmore, Updated Global Estimates of Respiratory Mortality in Adults ≥30 Years of Age Attributable to Long-Term Ozone Expos, 28 August 2017, https://doi.org/10.1289/EHP1390.

[41] Nan-Hung Hsieh, Chung-Min Liao, Fluctuations in air pollution give risk warning signals of asthma hospitalization, August 2013, https://doi.org/10.1016/j.atmosenv.2013.04.043.

[42] S. Du, T. Li, and Shi-Jinn Horng,Time Series Forecasting Using Sequence-to-Sequence Deep Learning Framework, 02 May 2019, 10.1109/PAAP.2018.00037.

[43] T.Liu, A. K. H. Lau, K. Sandbrink, J. C. H. Fung,Time Series Forecasting of Air Quality Based On Regional Numerical Modeling in Hong Kong, 24 March 2018, https://doi.org/10.1002/2017JD028052.

[44] NICOLÒ BALDON, Time series Forecast of Call volume in Call Centre using Statistical and Machine Learning Methods, Sweden 2019, urn:nbn:se:kth:diva-265002.

[45] S. G. Gocheva-Ilieva, A. V. Ivanov, D. S.Voynikova, and D. T. Boyadzhiev, Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach, 25 September 2013, https://doi.org/10.1007/s00477-013-0800-4.

[46] G. Papacharalampous, H. Tyralis, and D. Koutsoyiannis, Univariate Time Series Forecasting of Temperature and Precipitation with a Focus on Machine Learning algorithms: a Multiple-Case Study from Greece, 29 November 2018, https://doi.org/10.1007/s11269-018-2155-6.