Research Article

Multi-layer long short-term memory (LSTM) prediction model on air pollution for Konya province

Yahya Kocak a,*, Murat Koklu b

aThe Graduate School of Natural and Applied Science, Selcuk University, 42250 Konya, Turkey
bDepartment of Computer Engineering, Selcuk University, 42250 Konya, Turkey

ARTICLE INFO

Article history:
Received 21 November 2022
Accepted 27 December 2022

Keywords:
Air pollution
Air quality
Artificial neural network
LSTM
Multi-layer
Prediction

ABSTRACT

One of the main problems of the developing and changing world is air pollution. In addition to human causes such as population growth, increase in the number of vehicles producing exhaust emissions in line with the population, development of industry, natural causes such as forest fires, volcano eruptions and dust storms also play a role in increasing air pollution. Air pollution has become a bigger problem that reduces the quality of life of living beings and causes various lung and heart diseases due to reasons such as the growing proximity of settlements to industrial zones due to population growth, the increase in the number of individual vehicles, and zoning works carried out by ignoring air quality. Both international organizations and local authorities take various measures to control and prevent air pollution. In Turkey, necessary legal arrangements have been made within the scope of these measures and air quality monitoring stations have been established. The task of these stations is to measure pollutants such as PM10, CO, SO2 together with meteorological data such as air temperature, humidity, wind speed and direction. In this study, a prediction model for the future concentrations of PM10, CO and SO2 pollutants using the measurement data from three different air quality monitoring stations in Konya between January 2020 and January 2021 was realized with a multi-layer Long Short Term Memory (LSTM) artificial neural network. The Root Mean Square Deviation (RMSE) and Mean Absolute Percentage Error (MAPE) methods was used to calculate the performance of the study. As a result of the study, it is observed that the multi-layer LSTM architecture is more successful than the single-layer architecture.

This is an open access article under the CC BY-SA 4.0 license. (https://creativecommons.org/licenses/by-sa/4.0/)

1. Introduction

Air pollution is one of the leading factors that directly affect human health and reduce the habitability of the world ecosystem. It can cause many diseases, especially respiratory diseases and heart diseases, leading to a decrease in the quality of life and subsequently deaths. In 2019, it is estimated that nearly 307,000 premature deaths in Europe were caused by particles smaller than 10 micrometers and particles smaller than 2.5 micrometers, expressed as PM10 and PM2.5, and nearly 40,000 by nitrogen dioxide (NO2) [1]. Therefore, air pollution is an important issue that needs to be controlled and monitored [32, 33, 34, 35].

The air quality guidelines published by the World Health Organization (WHO) to evaluate air pollution in terms of public health and to determine a global road map have limit values for particulate matter (PM10 - PM2.5), nitrogen dioxide (NO2), carbon monoxide (CO), sulfur dioxide (SO2) and ozone (O3) pollutants. These guidelines were first published in 1987 and last updated in 2021. With the last amendment, the 24-hour limit value for PM10 pollutant was changed from 50 µg to 45 µg, the 24-hour limit value for NO2 pollutant was changed from undefined to 25 µg, the 24-hour limit value for SO2 pollutant was changed from undefined to 25 µg, the 24-hour limit value for CO pollutant was changed from undefined to 4 µg [2, 4, 36, 37].

In order to minimize air pollution, a number of legal regulations have been introduced in different countries for
years [4]. In order for the legal regulations to be implemented in the field, the factors that constitute the source of air pollution should be kept under control and necessary measures should be taken in advance. For this reason, air quality monitoring stations are established and the level of air pollutants is continuously monitored. At this point, the most important requirement of the decision support systems needed is a prediction model for the future [5]. Methods that can be used in this context include methods based on statistical calculations, rule-based classification methods and artificial neural networks [13, 27, 29, 42, 44]. The fact that deep learning techniques based on artificial neural networks are frequently used as classification and prediction models brings these methods to the forefront in new studies. [3, 6, 22, 31].

In this study, a prediction model is developed with long short-term memory (LSTM) artificial neural network architecture. In this model, 365 days of data from three different air quality monitoring stations in Konya province of Turkey, namely Erenköy, Karatay and Sakarya, between January 2020 and January 2021 are used. The performance of the model for PM10 pollutant was calculated separately for three different datasets using the Root Mean Square Deviation (RMSE) and Mean Absolute Percentage Error (MAPE) methods.

2. Related works

In recent years, studies on air pollution prediction have been carried out with different approaches, and data sets with different parameters from different regions have been used in these studies. Apart from digital sensor data, many studies have been carried out with different data such as traffic density, terrain and building data [11, 12, 14, 15, 16, 18, 20, 21, 23, 24, 26, 30]. The recent use of recurrent neural networks as a prediction method in similar data is also observed in this field.

46 scientific articles were analyzed in the research conducted in 2018. It was found that two types of machine learning approaches were used to predict air quality. In the first approach, ensemble learning and regression algorithms were used. With the algorithms used, it is aimed to predict the concentration of mass pollutants. It is observed that they provide an excellent balance between the interpretability and persistence of the model. In the second approach, the prediction problems are addressed using NN and SVM techniques [9]. Forecasting prioritizes accuracy over interpretability. It is explained why such powerful algorithms, which are also considered as a black box, are preferred. The study shows that the precision of the forecast tends to be lower and more variable than the accuracy of the prediction. It justifies the need to apply more computationally demanding (e.g. deep learning) methods to overcome the complexity of predicting the value of pollutants hours or days in advance. The study concludes that machine learning is a suitable method for predicting air pollution [24].

A study was conducted in Tarhan, the capital city of Iran, to identify prediction models to determine air pollution based on PM10 and PM2.5 pollution concentrations. Data on day of the week, month of the year, topography, meteorology and pollutant levels of the two nearest neighbors were used as input parameters and machine learning methods were used to predict air pollution. Using the genetic algorithm, day of the week, month of the year, topography data, wind direction, maximum temperature and pollutant ratio of the two nearest neighbors were determined as more effective parameters than other parameters in predicting air pollution. After investigating various methods for noise and error removal, the Savitzky-Golay filter was found to be more suitable than other methods such as wavelet analysis. A comparative study of machine learning methods including NARX, ANN, GWR and SVR was used for air pollution prediction and finally NARX was found to be optimal [23].

Yue-Shan et al. aimed to predict PM2.5 pollutant more successfully by using multiple LSTM models with different parameters collectively. The combined use of different learning models based on different features was combined in the final layer to obtain a more efficient prediction model [30].

Many air quality prediction models have been realized with artificial neural networks trained with measured foreign gas and particle amount data in the air. In a study conducted by Maleki et al. two groups of data sets, 95% training - 5% test and 90% training - 10% test, were obtained from foreign gas and particle concentration data from four measurement stations in Ahvaz region of Iran. Then, ANN based prediction was performed with these data sets and the correlation coefficient of the predicted values and measured values were calculated and compared with the Root Mean Square Deviation (RMSE) method. As a result of this study, it was concluded that the prediction process with the group containing 95% training data was more successful [7].

Xu et al. argued that using a single prediction model cannot reflect an intuitive result, which will lead to low prediction accuracy. As a solution to this problem, they suggested that using more than one prediction model together would be a solution to this problem. They observed that the combination of CEEMDAN-CNN-LSTM produces more successful results. By using particle swarm optimization to determine the parameters of this combination, they contributed to the model to produce more successful results [5].

In another study by Qi et al. using convolutional neural networks, they also used urban dynamics in the training
of the CNN-LSTM model called Deep-AIR. They obtained a more successful model by calculating the temporal correlations of air pollutants with data such as traffic speed and density, height and density of buildings on the streets [10].

3. Material and methods

3.1. LSTM

Recurrent neural network (RNN) is frequently used in time series problems, but in some cases it falls short. The main reason for the low performance is the Vanishing Gradient problem. To define this problem; the activation function does not cause a significant change in the activation function even when the input is reduced to a certain range by the activation functions. For this reason, the layer cannot learn enough and causes forgetting. The LSTM architecture developed by Sepp and Jürgen incorporates gates to solve this problem [28, 38, 39, 40, 41]. First of all, the weight of the input is accepted as 0 and a value between 0 and 1 is generated by passing it through a sigmoid function. A value close to 0 indicates that the information is unimportant and should be forgotten, while a value close to 1 indicates that the information is important and should be transferred. As a result, important information is transferred and unimportant information is forgotten and a more successful model is obtained. Therefore, LSTM can produce a more successful result in time series. The method based on the principle that the LSTM algorithm is back-propagation, i.e. bidirectional, is called Bi-LSTM [17, 19, 25].

3.2. Data specifications

In this study, data obtained from 3 different air quality monitoring stations located at different locations in Konya province of Turkey were used. These stations are located in Karatay, Erenkoy and Sakarya regions. These stations were selected because each station has different environmental factors. The fact that the station in Karatay is close to industrial zones, Sakarya station is close to settlements with high-rise buildings and dense population, and Erenkoy station is located in a single-storey and low population density area increases the diversity of the data. The study was conducted with three different data sets of 730 days and 17544 hours between January 1, 2020 and January 1, 2022. These data were obtained from the databank of the Continuous Monitoring Center of the Ministry of Environment, Urbanization and Climate Change of the Republic of Turkey [43]. The preliminary view of the data is as shown in Table 1. The content of the data includes date, PM10 (µg), SO2 (µg), NO2 (µg), NO (µg), NOx (µg), O3 (µg), CO (µg), temperature (oC), wind speed (km/h), wind direction (rad), relative humidity (%) parameters.

3.3. Data preprocessing

First of all, it is necessary to examine how healthy the data are. Because these data reflect the values measured by the sensors in the air quality monitoring stations. For example, if the sensor malfunctions, a different value may be read than it should be, or there may be no data in cases such as power outages and malfunctions. For this reason, the summary information of these data will be displayed first and it will be possible to decide which preprocessing will be carried out in the light of data such as number of data, average value, standard deviation, minimum value, maximum value. Summary information of the data of Karatay station is as shown in Table 2.

As seen in Table 2, some pollutants have negative values and unusually high values. The presence of these data in the data set will seriously affect the success rate of the model. For this reason, interval analysis of the data was performed with Quantile Regression. Figure 1 shows the analysis performed for PM10 pollutant.

<table>
<thead>
<tr>
<th>date</th>
<th>pm10</th>
<th>so2</th>
<th>no</th>
<th>no2</th>
<th>nox</th>
<th>o3</th>
<th>co</th>
<th>temp</th>
<th>wind_speed</th>
<th>wind_dir</th>
<th>rh</th>
</tr>
</thead>
<tbody>
<tr>
<td>01-01-2020 00:00</td>
<td>48.9</td>
<td>12.405</td>
<td>35.612</td>
<td>75.719</td>
<td>40.107</td>
<td>130.8</td>
<td>4.104</td>
<td>2.04</td>
<td>0.63</td>
<td>210.9</td>
<td>86.2</td>
</tr>
<tr>
<td>01-01-2020 01:00</td>
<td>58.4</td>
<td>11.51</td>
<td>24.852</td>
<td>65.122</td>
<td>40.27</td>
<td>261.26</td>
<td>4.114</td>
<td>2.25</td>
<td>0.56</td>
<td>114.6</td>
<td>85</td>
</tr>
<tr>
<td>01-01-2020 02:00</td>
<td>88.1</td>
<td>12.16</td>
<td>28.234</td>
<td>69.188</td>
<td>40.954</td>
<td>314.2</td>
<td>4.263</td>
<td>2.16</td>
<td>0.65</td>
<td>197.4</td>
<td>82.8</td>
</tr>
</tbody>
</table>
Table 2. Summary information of Karatay station

<table>
<thead>
<tr>
<th></th>
<th>pm10</th>
<th>so2</th>
<th>no</th>
<th>no2</th>
<th>nox</th>
<th>o3</th>
<th>co</th>
<th>temp</th>
<th>wind_speed</th>
<th>wind_dir</th>
<th>rh</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>17544</td>
<td>17544</td>
<td>17544</td>
<td>17544</td>
<td>17544</td>
<td>17544</td>
<td>17544</td>
<td>17544</td>
<td>17544</td>
<td>17544</td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>48.67</td>
<td>8.26</td>
<td>20.72</td>
<td>45.76</td>
<td>25.29</td>
<td>116.11</td>
<td>18.54</td>
<td>13.27</td>
<td>0.01</td>
<td>8.24</td>
<td>47.06</td>
</tr>
<tr>
<td>std</td>
<td>104.24</td>
<td>22.52</td>
<td>55.91</td>
<td>62.55</td>
<td>16.75</td>
<td>179.66</td>
<td>14.43</td>
<td>9.62</td>
<td>0.08</td>
<td>48.75</td>
<td>21.01</td>
</tr>
<tr>
<td>min</td>
<td>0</td>
<td>-13.34</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1785.59</td>
<td>0</td>
<td>-12.94</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>max</td>
<td>1000</td>
<td>1304.82</td>
<td>823.33</td>
<td>913.86</td>
<td>263.1</td>
<td>60625.8</td>
<td>262.26</td>
<td>37.12</td>
<td>1.52</td>
<td>298</td>
<td>91.4</td>
</tr>
</tbody>
</table>

Figure 1. Analysis result for PM10 pollutant

When the analysis in Figure 1 is examined, the data above and below a certain threshold value should be removed from the data set. After these data are removed, a healthier model training will be realized.

Another important point is the determination of the relationship between features. The relationship between traits plays an important role in determining the parameters to be used in the estimation of values. For this reason, the correlation matrix of the data set was calculated. The correlation matrix output is given in Figure 2.

There is a linear relationship between the values close to the yellow color in the correlation matrix. It is possible to say that when one of these values increases, the other also increases. There is an inverse relationship between the values close to the dark blue color. It is possible to say that while one of these values increases, the other decreases. For example, an increase in NO value allows us to predict an increase in NO2 and O3 values. Likewise, we can say that NOx value increases inversely with wind speed. Likewise, there is an inverse relationship between air temperature and all pollutants. The main reason for this is that there is a relationship with the increased use of fossil fuels when the weather is colder.

3.4. Experimental design

Following the preprocessing stage, the data were normalized using the MinMaxScaler method. Since outliers are removed by Quantile regression, the MinMaxScaler method performs better. Since prediction will be performed in this study, supervised learning was chosen as the learning technique for the data. The advantage of the supervised learning technique is that it provides an evidence-based learning process. The window size was set to 4. The data set is divided into 70% learning and 30% test data. A 3-layer LSTM architecture was defined. In this architecture, 50 neurons are defined for each LSTM layer.

In deep learning applications, optimization is a process that involves finding the optimal value in a given problem. There are various optimization algorithms that are commonly used in this type of system. Some of these include adagrad, adadelta, adamax, and stochastic gradient descent.
Dropout technique was used to increase the prediction accuracy by preventing overfitting during learning. Dropout aims to prevent overlearning by randomly canceling the operation of some neurons. Usually, a value of 0.5, which means 50%, is used. In this study, a dropout value of 0.4, which can be expressed as 40%, was used.

The training of the model was performed in 50 epochs and this process was also applied for the data of three different stations. As a result of the model, it is aimed to predict the PM10 concentration.

4. Experimental Results

The aim of this paper is to estimate the PM$_{10}$ pollutant using air pollution data from Konya region using a multilayer LSTM architecture. Within the scope of the study, training procedures were performed for 3 different stations and error values were calculated.

RMSE and MAPE methods were used as error calculation methods. Root Mean Square Deviation (RMSE) produces an absolute number of how much the predicted values differ from the actual number. While no single result is completely meaningful, it is a benchmark for comparison with other model results and can be used to select the best model. Root Mean Square Deviation (RMSE) is calculated by taking the square root of the MSE value. RMSE is used more often because MSE may be too large to be used in model comparisons. Therefore, the MSE is calculated as the square of the error and is easier to interpret. However, the MSE is very sensitive to unusual values.

Mean Absolute Percentage Error (MAPE) is the mean absolute difference between the true value and the predicted value expressed as a percentage. This method also enables the model to produce a comparable success value.

The result data for the air quality monitoring station located in Karatay region are given in Table 3.

The actual value and predicted value graph as a result of training for Karatay region is given in Figure 3.

### Table 3. Result information of Karatay station

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>RMSE for single-layer</th>
<th>MAPE for single-layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Window Size</td>
<td>4</td>
<td>15.14489</td>
<td>19.15937</td>
</tr>
<tr>
<td>LSTM units</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dropout value</td>
<td>0.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dense units</td>
<td>1</td>
<td>11.60648</td>
<td>12.64851</td>
</tr>
<tr>
<td>Epoch</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Batch Size</td>
<td>32</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The aim of this paper is to estimate the PM$_{10}$ pollutant using air pollution data from Konya region using a multilayer LSTM architecture. Within the scope of the study, training procedures were performed for 3 different stations and error values were calculated.
The result data for the air quality monitoring station located in Ereğli region are given in Table 4.

Table 4. Result information of Ereğli station

<table>
<thead>
<tr>
<th>Window Size</th>
<th>4</th>
<th>RMSE for single-layer = 19.687252</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM units</td>
<td>50</td>
<td>MAPE for single-layer = 26.852158</td>
</tr>
<tr>
<td>Dropout value</td>
<td>0.4</td>
<td>RMSE for multi-layer = 14.490823</td>
</tr>
<tr>
<td>Dense units</td>
<td>1</td>
<td>MAPE for single-layer = 19.748219</td>
</tr>
<tr>
<td>Epoch</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Batch Size</td>
<td>32</td>
<td></td>
</tr>
</tbody>
</table>

The actual value and predicted value graph for Ereğli region is given in Figure 4.

Figure 4. Graph between actual and estimated values for Ereğli region

The result data for the air quality monitoring station located in Sakarya region are given in Table 5.

Table 5. Result information of Sakarya station

<table>
<thead>
<tr>
<th>Window Size</th>
<th>4</th>
<th>RMSE for single-layer = 19.687252</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM units</td>
<td>50</td>
<td>MAPE for single-layer = 22.298347</td>
</tr>
<tr>
<td>Dropout value</td>
<td>0.4</td>
<td>RMSE for multi-layer = 14.490823</td>
</tr>
<tr>
<td>Dense units</td>
<td>1</td>
<td>MAPE for single-layer = 16.981358</td>
</tr>
<tr>
<td>Epoch</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>Batch Size</td>
<td>32</td>
<td></td>
</tr>
</tbody>
</table>

The actual value and predicted value graph for Sakarya region is given in Figure 5.

Figure 5. Graph between actual and estimated values for Sakarya region

5. Conclusion and discussion

In this study, by comparing the RMSE and MAPE values, we show that the training model using the multi-layer LSTM architecture has a better prediction performance than the single-layer LSTM architecture.

The data of 3 different regions of Konya province were used in the study and the error comparisons of the data sets used are given in Table 2, Table 3 and Table 4. When the numerical error data obtained in the study are compared, both the RMSE value and the MAPE value are lower, indicating that the multilayer LSTM architecture is a more accurate model for air quality prediction.

When the numerical error data obtained in the study are compared, both the RMSE value and the MAPE value are lower, indicating that the multilayer LSTM architecture is a more accurate model for air quality prediction.

More accurate results can be achieved with different data sets, different parameters, different optimization and normalization processes. At the same time, a more
performant prediction model can be obtained by using different deep learning techniques.

References


