Chapter

A Layered Recurrent Neural Network for Imputing Air Pollutants Missing Data and Prediction of NO_2, O_3, PM_{10}, and $PM_{2.5}$

Hamza Turabieh, Alaa Sheta, Malik Braik and Elvira Kovač-Andrić

Abstract

To fulfill the national air quality standards, many countries have created emissions monitoring strategies on air quality. Nowadays, policymakers and air quality executives depend on scientific computation and prediction models to monitor that cause air pollution, especially in industrial cities. Air pollution is considered one of the primary problems that could cause many human health problems such as asthma, damage to lungs, and even death. In this study, we present investigated development forecasting models for air pollutant attributes including Particulate Matters ($PM_{2.5}$, PM_{10}), ground-level Ozone (O_3), and Nitrogen Oxides (NO_2). The dataset used was collected from Dubrovnik city, which is located in the east of Croatia. The collected data has missing values. Therefore, we suggested the use of a Layered Recurrent Neural Network (L-RNN) to impute the missing value(s) of air pollutant attributes then build forecasting models. We adopted four regression models to forecast air pollutant attributes, which are: Multiple Linear Regression (MLR), Decision Tree Regression (DTR), Artificial Neural Network (ANN) and L-RNN. The obtained results show that the proposed method enhances the overall performance of other forecasting models.

Keywords: imputing missing data, air pollutants, prediction, layered recurrent neural network

1. Introduction

Air quality monitoring and management have drawn much attention in recent years and attracted great attention from the public. Air pollution poses serious problems and infection for living organisms and environmental risks [1]. Harmful emission of industrial waste on air is one of the common environmental influences that disturb the air quality specifications and the national economy [2]. Significant publications have shown that air pollution has harmful effects on human health [3]. Air pollution affects the living organisms by producing impacts on cardiac, vascular, pulmonary, and neurological systems [4]. For example, air pollution in the City

of New York causes the death of more than 3000 people and causes hospitalization of 200 persons [5]. It was found that many of these reported incidences were caused by the exposure to $PM_{2.5}$ and other pollutant attributes [6]. In 2010, it was estimated that ambient particulate matter (PM) caused 3.2 million premature deaths [7]. Moreover, several-analysis and research papers highlight that there is an exponential relationship between PM values and cardiovascular disease, and significant relation between NO_2 concentrations and cardiovascular disease [8, 9].

Air pollution arises from many sources such as vehicle fumes, agricultural, industrial, and natural sources like volcanoes [10]. Common air pollutants are classified into two groups: trace gases such as carbon monoxide (CO), nitrogen dioxide (NO_2), ground-level ozone (O_3), and sulfur dioxide (SO_2) or particulate matter ($PM_{2.5}$) or (PM_{10}) in aerodynamic diameter [11]. Tropospheric ground-level ozone (O_3) is a secondary factor that can damage human health and ecosystem [12–41]. O_3 concentration is one of the most serious oxidant factors that are harmful to human skin and lung tissues when inhaled [15, 16]. Several side effects impair pulmonary function and cause respiratory symptoms such as headache, weight loss, cough, shortness of breath, hoarseness, and pain while breathing [17]. Moreover, several epidemiological research studies focus on the relation between O_3 pollution and mortality [18].

1.1 Challenges

Air pollution monitoring and control is a major global challenge [19, 20]. To develop and train air quality prediction models, meteorological data for the investigated area should be collected and used. This data mostly consists of physical parameters that include temperature, dew point, wind direction, wind speed, cloud cover, cloud layer(s), ceiling height, visibility, current weather, amount of precipitation, and many more [21, 42]. These attributes greatly influence the concentration of pollutants in the area of interest.

Recently, cities are exposed to air pollutants either indoors or outdoors [22]. Several monitoring stations (i.e., sensors) are used to monitor the air quality by collecting data from different locations inside cities. These stations are used to collect data for gases or particulate matter in an accurate manner [23]. These sensors can be categorized as wired or wireless sensors. Wired sensors need great efforts for deployment and maintenance. Wired sensors can be easily breakdown due to several reasons (e.g., environment close to a volcano, where the hot gases and steams can damage a wired network easily [24, 43]). Wireless sensors still in an early stage. However, they show a great performance compared to wired sensors either in deployment or maintenance. Both types of sensors send the collected data to a central station for further processing. However, sometimes the process of collecting data suffers from different problems such as power failure, sensor fault, man-made error in measurements, and many others. **Figure 1** depicts the process of collecting data from different sensors. For example, if the gas sensor (i.e., O_3) does not work accurately, the collected data will not be complete and accurate. As a result, the air quality of the prediction model may not be accurate if the percentage of missing data is high.

Missing data cause serious problems for developing prediction models. The presence of missing data could severely reduce the quality of air quality prediction models. To solve this problem, we may either remove the missing data or imputing it. Removing the missing data may reduce the application performance [25], while imputing missing data may enhance the overall performance and without losing the collected data. Many methods exist to impute the missing data. Researchers either applied simple methods such as average value or complex methods such as machine

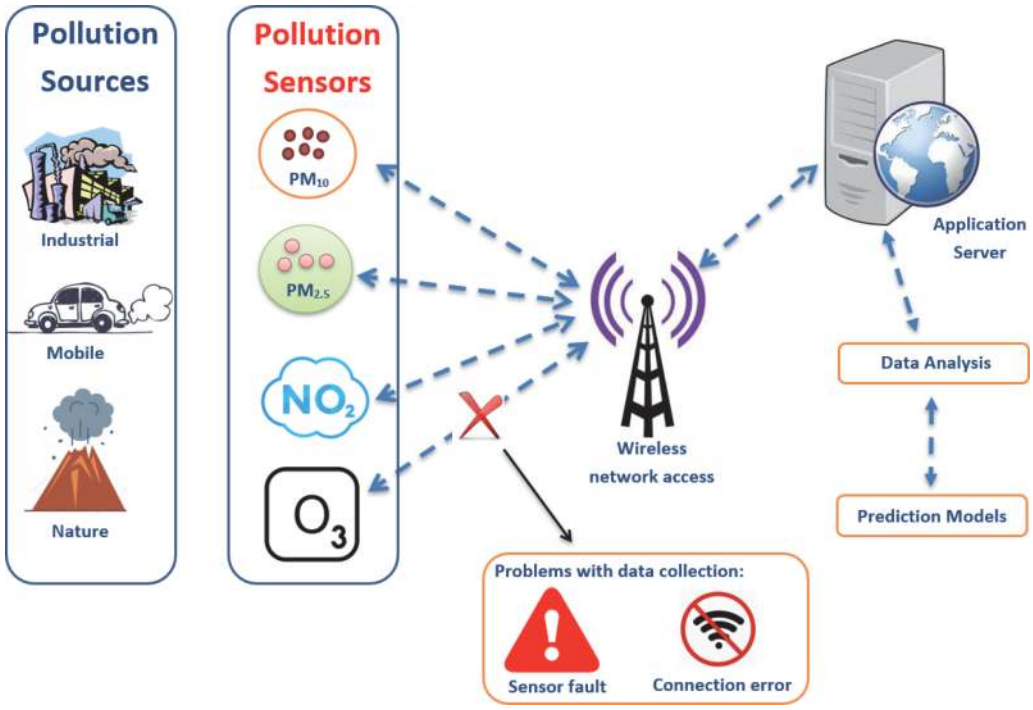

Figure 1.
The process of collecting data using air pollutants sensors.

learning methods to impute missing data [25]. Imputing missing values based on average is not accurate compared to the other one.

The main goal of this study is to propose a hybrid model that can predict the daily average of the concentration of air pollutants based on missing data imputation. The proposed model is a machine learning approach that can enhance the performance of monitoring systems of air pollution inside cities. Layered Recurrent Neural Network (L-RNN) [26] was successfully used to solve a variety of state-of-the-art applications such as detection of heart failure [27] and time-series data classification [28]. L-RNNs for missing data were explored earlier to handle the missing data problem [25, 29]. In this research, we first explore the use of L-RNN for imputing the missing data collected from Dubrovnik city that is in the east of Croatia. In the second step, we develop a series of models for predicting NO_2, O_3, PM_{10}, and $PM_{2.5}$ using the machine learning model (i.e., MLR, DTR, ANN, and L-RNN).

The rest of this chapter is organized as follows: In Section 2, the related works of air pollution as well as the literature of missing data is presented. Section 3 describes the proposed methodology. Section 4 presents predictive models using machine learning concepts. Section 5 presents the data collection process. The evaluation criteria used in this chapter are presented in Section 6. Section 7 presents the experimental setup used in this paper. Finally, a conclusion of the work is presented in Section 8.

2. Related works

2.1 Imputation vs. removing data

One of the most common problems in the process of developing prediction models is the Data Cleaning/Exploratory Analysis. This phase becomes a challenge

when missing values are in presence. In general, there is no fundamental method to deal with missing data. Missing data problem occurs if no value(s) is assigned while collecting data. In general, the missing data are presented by different symbols such as ?, N/A, or −. There are two methods adopted in the literature to handle missing data. They are:

2.1.1 Deletion

Several researchers remove the missing data from the collected dataset if the percentage of the missing data is less than 5%. However, if the percentage of missing data is greater than 5%, the dataset should be examined carefully [30]. Many approaches have been investigated by researchers to solve the missing data problem. For example, the data list wise deletion method removes the missing data or incomplete data from the collected dataset. This method works fine if the percentage of missing data is very small and does not affect the overall accuracy [31]. The pairwise data deletion method keeps the missing data and tries to reduce the loss that occurs in the list wise deletion method. However, deleting missing values is an acceptable approach for some applications. Mary and Arockiam [32] investigated the missing data as a case study of air pollution. They proposed an ST-correlated proximate approach to impute the incomplete dataset for the air pollution system. The authors compared the obtained results of the proposed approach with different statistical methods. Sta [33] investigated the process of collecting data for modern urban cities. The author proposed a framework to cluster the collected data into three clusters: complete, ambiguous, and missing data. The author imputed the missing data and enhanced the overall performance of the proposed system. Xiaodong et al. [34] proposed a Hot Deck imputation approach that imputes the incomplete records (missing data) using the similarity between complete and incomplete data.

2.1.2 Imputation

In statistics, imputation is defined as the process of substituting missing data with swapped values. Unit imputation is used when we replace a single data point while the replacement of a component of a data point, is called, item imputation. Imputation is considered a successful solution to avoid difficulties associated with list wise deletion of missing values. Suhani et al. [35] proposed a machine learning approach based on the fuzzy kNN technique to impute the missing data for a selected case from the medical field. The authors ignore the missing data whose entropy value is less than a predetermined value and recover the incomplete data that are higher than the predetermined value based on a fuzzy kNN algorithm. Chen et al. [36] applied a machine learning approach based on a convolutional neural network to impute the missing data for a real medical dataset. The authors improve the overall performance after imputing missing data. Turabieh et al. [25] proposed a dynamic model based on deep learning neural networks for missing data imputation. The authors showed that the proposed model improves the overall performance of medical applications after imputing missing data.

2.2 Air pollution prediction

Air pollution is a serious problem that negatively affects human health, environment, and climate. Governments and organizations published several initiatives to reduce the concentrations of air pollutants, but high levels of concentrations of air pollutants still exist. As a result, monitoring the concentrations of air pollution is

needed. Air monitoring consists of several steps; 1) Monitoring sites based on wired or wireless sensors, 2) collecting data that should be accurate and complete, 3) data analysis using predictive models based on machine learning to predict and analyze the collected data, and, 4) the final step is making decisions to reduce the concentrations of air pollution. This process should be performed correctly to ensure that the concentration of air pollution is under control.

Different types of machine learning methods have been used to predict the concentrations of air pollutant indicators by many researchers. For example, Perez and Gramsch [37] applied a feed-forward neural network to predict the concentration of $PM_{2.5}$ and PM_{10} in Santiago, Chile. The obtained accurate results show that the proposed approach enhances the prediction of $PM_{2.5}$ and PM_{10}. Lana et al. [38] employed regression models to predict several air pollutants such as CO, NO, NO_2, O_3 and PM_{10}) for the city of Madrid (Spain). The obtained results explore the importance of reducing air pollutants in the city of Madrid. Kamińska [39] employed an ensemble learning method based on random forests to model the relationship between the concentrations of air pollutants and nine variables describing meteorological conditions, temporal conditions, and traffic flow. The collected data was for 2 years 2015 and 2016 for WrocÅ‚aw city. The data consists of hourly values of wind speed, wind direction, temperature, air pressure and relative humidity, temporal variables, and traffic flow. The obtained results show that the season plays a vital role in the overall performance. Kamińska [40] proposed a probabilistic forecasting method to predict the concentrations of NO_2. The dataset represents the hourly values of the concentration of NO_2 wind speed, and traffic flow for the main intersection in Wrocław city. The obtained results show that wind speed plays a vital factor in the concentration of NO_2.

Shang et al. [41] employed a novel prediction method that hybridized the regression tree (CART) and ensemble extreme learning machine (EELM) methods to predict the hourly concentration of $PM_{2.5}$ air pollutant. The training dataset used in this research obtained from the meteorological data of Yancheng urban area, while the testing data (i.e., the air pollutant concentration) obtained from the City Monitoring Centre. The obtained results demonstrate the effectiveness of the proposed method to predict $PM_{2.5}$. A hybrid framework based on three different machine learning methods (i.e., genetic algorithm [GA], random forests [RFs], and backpropagation neural networks [BPNN]) proposed by Dotse et al. [44]. The proposed hybrid approach is used to predict daily PM_{10} in Brunei Darussalam. Sun and Sun [45] proposed a hybrid model to predict $PM_{2.5}$ in Baoding city in China, where a combination of three machine learning methods (i.e., principal component analysis [PCA], least squares support vector machine [LSSVM], and cuckoo search [CS]). The obtained results show that the PCA algorithm works as a feature selection algorithm that reduces the dimensionality of the input dataset while CS shows promising results to predict $PM_{2.5}$. The main shortfall of this work that is applicable for short-term $PM_{2.5}$ forecasting.

A dynamic fuzzy synthetic evaluation model for predicting the concentration of three air pollutants (i.e., of $PM_{2.5}$, PM_{10} and SO_2) in two cities from China have been proposed by Xu et al. [46]. The obtained results show that the proposed model can be employed to build a robust monitoring air quality system for early warning. A novel hybrid model based on extreme learning machine (ELM) is employed to predict the concentration level of PM_{10} for Beijing and Harbin cities in China by Luo et al. [47]. Aznarte [48] proposed an ELM approach that is optimized by cuckoo search (CS) to enhance the overall performance of ELM. A probabilistic forecasting approach is applied to predict NO_2 in Madrid city from Spain. Wang et al. [49] proposed a novel hybrid machine learning approach based on a decomposition method and extreme learning machine (ELM) that is optimized by differential

evolution (DE) to predict air pollutants in Beijing and Shanghai cities from China. Kumar and Goyal [50] applied Multiple Linear Regression (MLR) and Principal Component Regression (PCR) methods to predict several air pollutants in Delhi city from India. A MultiLayer Perceptron (MLP) neural network is adopted to predict PM_{10} in Delhi city from India by Aly et al. [51]. The authors also applied two algorithms (i.e., Naïve Bayes [NB] and Support Vector Machine [SVM]) and the performance of MLP outperforms NB and SVM. Vibha and Satyendra [52] applied seven models of neural networks using Levenberg–Marquardt (LM) to predict the daily PM_{10} in two cities from India.

3. Methodology

The main purpose of this research is to evolve different machine learning methods to predict daily average air pollutant concentrations such as O_3, $PM_{2.5}$, PM_{10}, and NO_2 values given data with many missing values. The process consists of two phases: 1) imputing missing data based L-RNN model, and 2) development of predictive models using several machine learning algorithms which include LR, DTR, ANN, and L-RNN. Our proposed approach starts by collecting data from sensors. If the collected data suffer from missing data, an imputation process will start based on the L-RNN hat predicts the concentration of air pollutants. This process will be repeated until the collected data have no missing value(s). Once the collected dataset is complete, a machine learning model is selected to predict the daily average of air pollutant concentrations. The selected model is evaluated based on two evaluation criteria that are Root Mean Square Error (RMSE), and coefficient of determination (R^2). The proposed approach is depicted in **Figure 2**. The following subsections demonstrate the proposed approach.

3.1 L-RNN

A layered recurrent neural network is known as a neural network that has local feedback, which is particularly suited to predict the daily air pollutant attributes since it incorporates a time delay while training process through a feedback

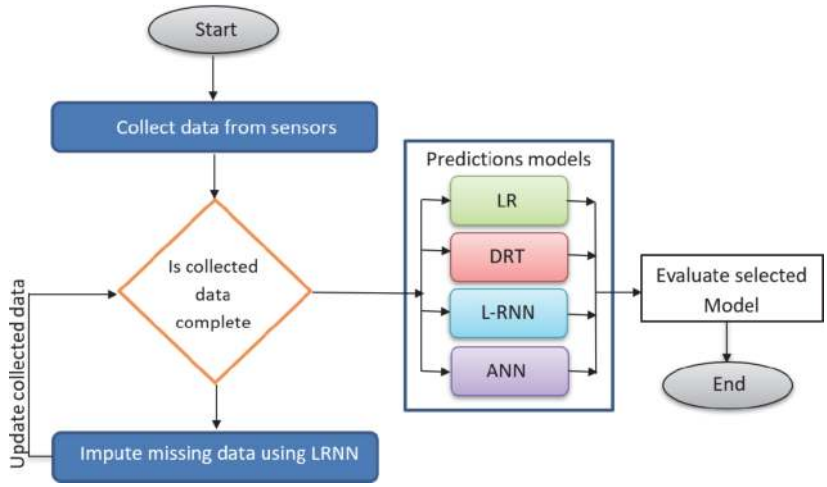

Figure 2.
The proposed approach.

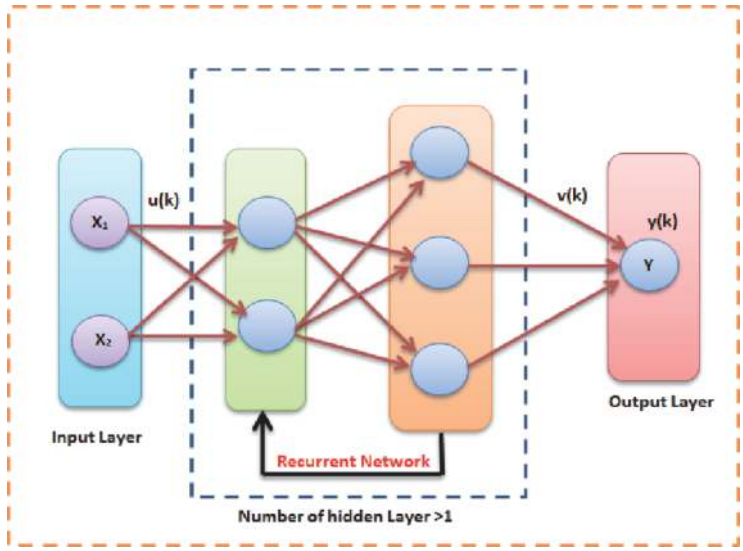

Figure 3.
Layer recurrent neural network (L-RNN).

connection between output layer and hidden layer(s). **Figure 3** demonstrates the connection feedback. In simple, during the training process, the output of the recurrent neural network is added to the output of the hidden layer. The result of summation is employed as an argument of the transfer function to gain the output in the succeeding iteration. Eq.(1) demonstrates the output of the L-RNN, where $u(k)$ presents the input values for hidden layer, $v(k)$ presents the input values for output layer. $W_{u,i}$ and $W_{v,j}$ presents the weights between u and v, respectively. The final output $y(k)$ is obtained from Eq.(2), where $f()$ is a transfer function. In this work, we employed back-propagation through time in the training phase for the proposed L-RNN structure.

$$v(k+1) = \sum_{i=0}^{n} W_{u,i}(k)u(k) + \sum_{j=0}^{m} w_{v,j}(k)v(k) \tag{1}$$

$$y(k) = f(v(k)) \quad where: f(v(k)) = \frac{1}{1 + exp\,(-(v(k))} \tag{2}$$

3.2 Data imputation using L-RNN

To implement the data imputation process, we clustered the data into two groups 1) complete dataset [without a missing value(s)] and 2) incomplete dataset [with a missing value(s)]. A holdout method is used to train and test the L-RNN. The complete dataset is divided into three datasets: training dataset (70%), testing dataset (15%), and validation dataset (15%). While the incomplete dataset is used to simulate the trained L-RNN model to impute the missing value(s). This process will be repeated dynamically while receiving any records with a missing value(s).

The computational complexity of the model depends on the structure of the L-RNN and the number of missing data in the received record. The computational complexity will increase exponentially if the number of missing values increases. **Figure 4** illustrates the process used to impute the missing value(s) (i.e., the concentration value of O_3, NO_2, $PM_{2.5}$, PM_{10}).

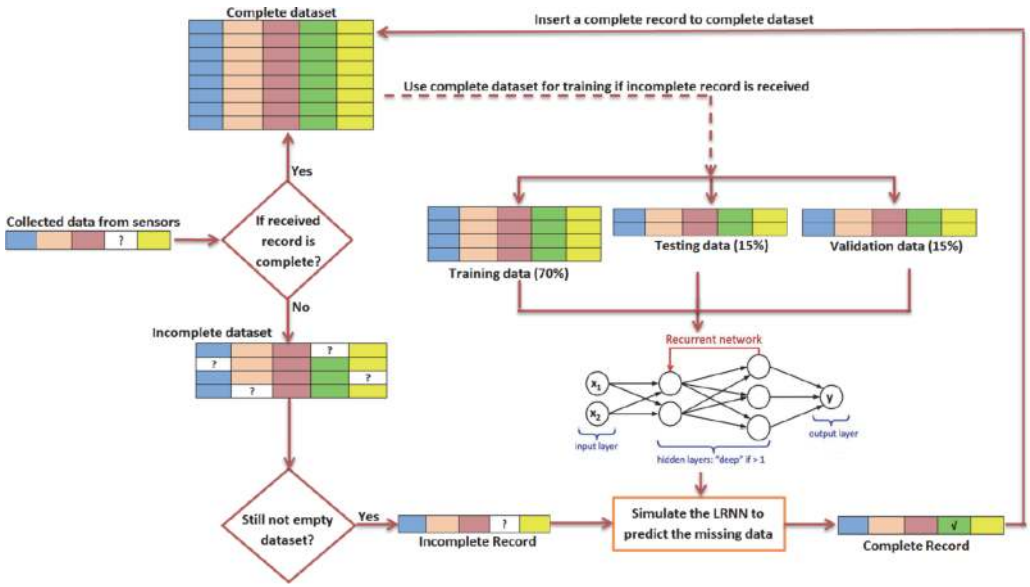

Figure 4.
Impute missing data approach.

4. Predictive models using machine learning

Several machine learning methods can be used to predict air quality. However, we have limited our research paper into four methods: MLR, DTR, ANN, and L-RNN. To avoid the over-fitting problem in the training process, we employed the k-fold cross-validation method with k-fold = 5. The following subsection explores each learning method in more detail.

4.1 Multiple linear regression

Linear regression (LR) is one of the most well-known algorithms in statistics and machine learning. The main idea of LR is to find the relationship between input and output numerical variables. There are several types of LR such as Simple linear regression, multiple linear regression, logistic regression, ordinal regression, Multinomial regression, and Discriminant Analysis. LR has been employed successfully in many areas as a machine learning algorithm [53, 54]. MLR is a classical statistical method that tries to find a relationship between complex input–output variables. In simple, MLR tries to find an approximation linear function between independent input variables and dependent output variable without loss of generality. Eq.(3) explores the regression line in MLR.

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_i x_i + \ldots + \beta_k x_k + \varepsilon \tag{3}$$

where y is dependent output variable, x_i is the i^{th} independent input variable, β_i is polynomial coefficients of x_i, k is the number of independent input variables, and $\varepsilon$is the possible variation form. Eq.(4) presents a compact version of Eq.(3).

$$y = X\beta + \varepsilon \tag{4}$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_n \end{bmatrix}, and \quad X = \begin{bmatrix} 1, x_{1,1}, \ldots, x_{1,k} \\ 1, x_{2,1}, \ldots, x_{2,k} \\ . \\ . \\ . \\ 1, x_{n,1}, \ldots, x_{n,k} \end{bmatrix}_{n \times (k+1)} \tag{5}$$

where n represents the number of samples, $x_{m,i}$ represents the value of i^{th} independent input variable in the m^{th} sample, and ε_i is the i^{th} residual error in the m^{th} sample. The coefficient vector β can be calculated based on the standard least-square method as shown in Eq.(6).

$$\beta = \left(X^T X\right)^{-1} X^T Y \tag{6}$$

Therefore, when the parameter vector β is known, the generated MLR model can predict the dependent output variable based on the independent input variable (s).

4.2 Decision tree regression

The DTR is employed in this chapter to predict the air pollutant attributes due to its ability to handle complex data and takes less training execution time compared to other prediction models. In simple, DTR uses if-then conditions to predict the appropriate output value(s) [55]. The DTR has three steps to predict the output value(s) as follows:

- **Step 1**: Determining the parameter settings for DTR such as: predicting accuracy, selecting splits, when to stop splitting, and selecting the optimal tree.

- **Step 2**: Selecting the splits to predict values of the continuous dependent variable, which usually measured with node impurity measure which provides an indication of the relative homogeneity of cases in the terminal nodes.

- **Step 3**: Determining when to stop the splitting which is related to the minimum number of nodes. Which means to select the best rightly-sized tree, which is called the optimum tree.

4.3 Artificial neural network

Artificial Neural Network (ANN) has been used widely in many forecasting applications due to its ability to handle complex data. Without having any information about the mathematical model that represents the relation between input and output variables, ANN can learn the learn hidden knowledge between input and output variables. In general, there are many kinds of ANNs such as Feedforward Neural Network (FFNN), Recurrent Neural Network (RNN), and Convolutional Neural Network (CNN) [56]. In this chapter, we adopted two types of neural networks based on a feed-forward network using the propagation

training method, which is: the standard neural network (ANN) and Layered Recurrent Neural Network (LRNN).

5. Data collection

The data set used in this research is collected from Dubrovnik city that is located in the east of Croatia. Dubrovnik city has a Mediterranean climate and has over 2600 hours of sunshine per year, which is considered the sunniest place in Croatia. In this dataset, the concentration of O_3 has been monitored with a commercial Teledyne API 400E UV photometric O_3 analyzer. While the concentration of NO_2 has been monitored with Teledyne API 200E chemiluminescent NO_2 analyzer. O_3 and NO_2 concentrations were measured every minute and the output signals were stored in a datalogger. The collected data are validated and averaged. The concentration of PM_{10}, and $PM_{2.5}$ have been monitored with the GRIMM model EDM 180. Samples of PM particles were collected by gravimetric methods throughout the day to obtain 24–hour averages of concentrations. All instruments are regularly maintained and calibrated. Meteorological data were obtained from the Meteorological and Hydrological Services of Croatia. The dataset is collected during the 2015 and 2016.

Table 1 shows the number of records in each dataset used in this paper. For example, the O_3 dataset has 699 total records, where 200 records (28.80%) are incomplete. **Figure 5b** demonstrates the missing data pattern for each dataset, where the x-axis presents the 24-hours (i.e., input variables), while the y-axis presents the observations during the 2 years. **Figure 5a** shows that there is a missing data in the second year for NO2 dataset, where NO_2 sensors do not work. Since the missing data are higher than 5%, we examined the collected data carefully to maintain the performance of air quality prediction systems. As a result, imputing missing data are needed.

6. Evaluation criteria

In this research, we employed two different evaluation criteria: Root Mean Square Error (RMSE), and coefficient of determination (R^2), defined below.

$$RMSE = \sqrt{\frac{1}{s}\sum_{i=1}^{s}\left(y_{i\,predicted} - y_{i\,observed}\right)^2} \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{s}\left(y_{i\,predicted} - y_{i\,observed}\right)^2}{\sum_{i=1}^{s}\left(y_{i\,observed} - \hat{y}_{i\,observed}\right)^2} \tag{8}$$

Dataset	$PM_{2.5}$	PM_{10}	O_3	NO_2
InComplete	179	179	200	270
Complete	551	551	499	461
Percentage of missing data %	16.96	17.08	20.26	28.80
Total number of records	730	731	699	731

Table 1.
Number of samples in each dataset.

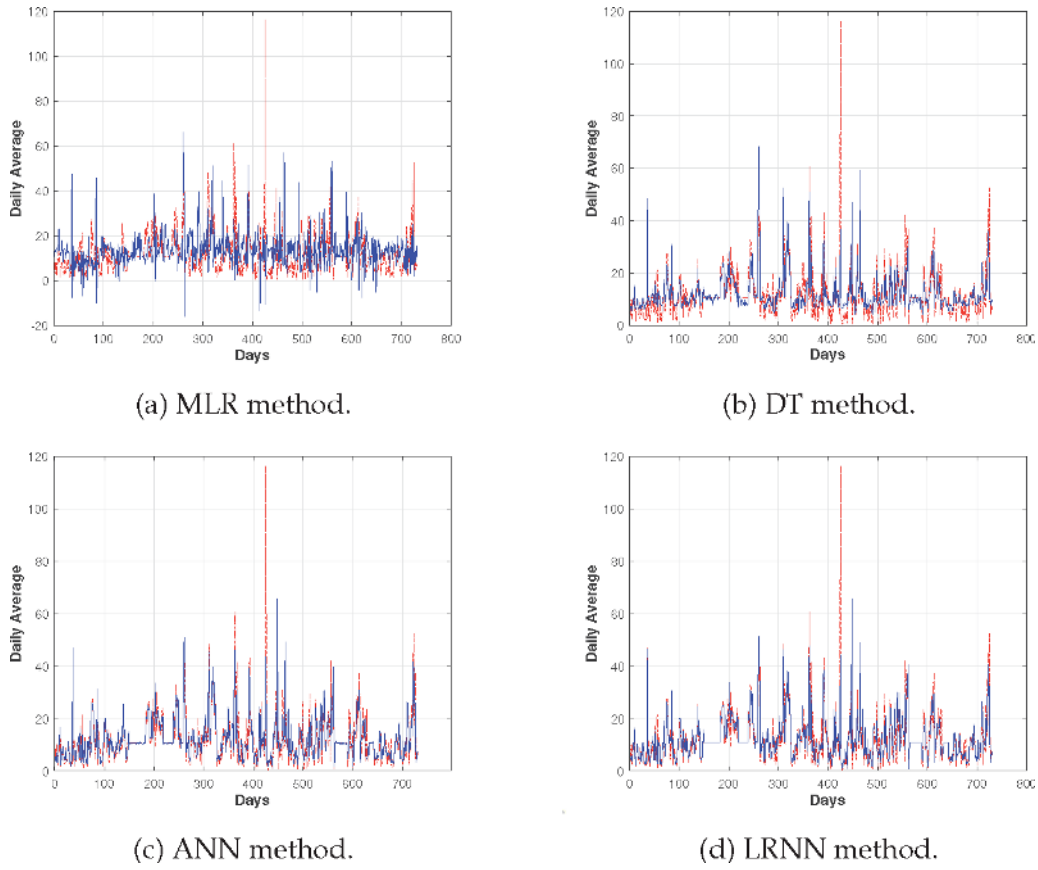

Figure 5.
Actual (–) and predicted (–.) values for No_2 using all regression methods for NO_2 dataset.

where $y_{i\,observed}$ and $y_{i\,predicted}$ denote the actual and predicted values of air pollution concentrations, respectively, s represents the number of instances and $\hat{y}_{i\,observed}$ stands for the average of the actual values of the air pollution concentrations.

Eqs.(7) and (8) show the evaluation process for each criteria. The minimum value of RMSE means better forecasting, while the maximum value of R^2 means better forecasting.

7. Experimental results

In this work, two different types of experiments were performed to develop a prediction model for pollutant parameters with missing data. They are: (i) removing missing or incomplete records, and (ii) imputing the missing data. Four regression models were employed in this work (i.e., MLR, DT, ANN, and LRNN). All experiments were performed using MATLAB-R2019b environment. The following subsections discussed the obtained results.

7.1 Results without imputing missing data

The first experiments that we employed in this chapter are based on removing all the missing data (i.e., records). **Table 2** shows the obtained results of four different regression models. The LRNN model outperforms other models in three

Regression model	NO_2		O_3		PM_{10}		$PM_{2.5}$	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
MLR	1.79	0.10	22.11	0.05	3.68	0.83	2.61	0.81
DT	1.73	0.16	20.44	0.19	3.68	0.83	2.57	0.82
LRNN	**0.26**	**0.85**	10.06	0.61	**0.30**	**0.88**	**2.39**	**0.90**
ANN	1.02	0.74	**8.39**	**0.76**	3.85	0.67	2.92	0.85

Table 2.
Results without imputing missing data.

datasets (i.e., NO_2, PM_{10}, and $PM_{2.5}$) based on RMSE and R^2 values. While ANN outperforms other models in O_3 dataset based on RMSE. The performance of the MLR method is the worst overall datasets.

7.2 Imputing data using LRNN

For imputing missing data, we employed L-RNN as a dynamic prediction model based on the current states of the collected data. In general, there are two different training algorithms for L-RNN: real-time recurrent learning, where a fixed set of weights recursively applied while training process and back-propagation through time, where the L-RNN structure altered between feed-forward and feedback structures. In this work, we used back-propagation through a time training process.

The parameters setting used, in this case, are shown in **Table 3**. A holdout method is employed to train the L-RNN based on the complete dataset, where 70% for training, 15% for validation, and 15% for testing. The reason for employing the holdout method is to reduce the complexity and execution time for the proposed imputing model. After imputing missing data, we employed a k-fold across-validation method in the training process for four machine learning methods (i.e., MLR, DTR, L-RNN, and ANN) with k-fold = 5 to evaluate the complete dataset.

7.3 Results after imputing missing data

7.3.1 MLR models

In this part, we employed MLR as a prediction model after imputing missing data. In Eq.(9), Eq.(10), Eq.(11), and Eq.(12) we show the MLR results for $PM_{2.5}$,

Parameters	Values
Number of iterations	1000
Number of neurons in the input layer	Number of input data
Number of neurons in the hidden layer	Number of input data /2
Number of neurons in the output layer	1

Table 3.
Parameter settings for the L-RNN model during imputing missing data.

Dataset	MLR results	
	RMSE	R^2
NO_2	1.61	0.79
O_3	1.78	0.82
PM_{10}	3.84	0.46
$PM_{2.5}$	1.76	0.62

Table 4.
MLR results after imputing missing data.

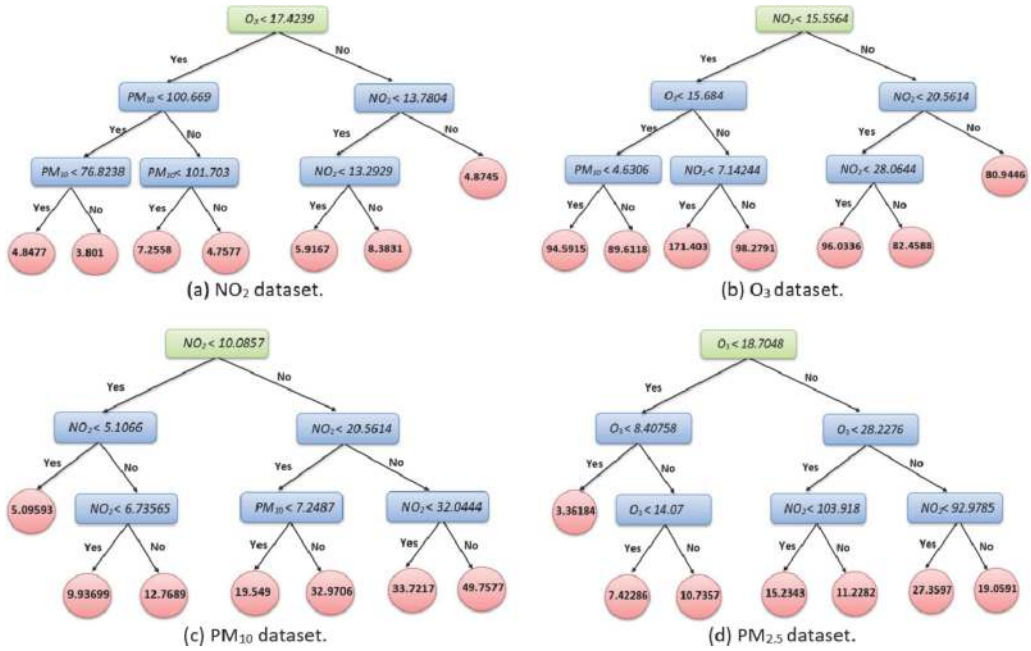

Figure 6.
Obtained tree for all datasets after imputing missing data.

PM_{10}, O_3, and NO_2, respectively. **Table 4** shows the obtained results of MLR method. The performance of of MLR is acceptable over all datasets.

$$PM_{2.5} = 4.4447 - 0.3753 \times NO_2 + 0.61551 \times PM_{10} - 0.025394 \times O_3 \quad (9)$$

$$PM_{10} = -3.8704 + 0.6183 \times NO_2 + 0.033796 \times O_3 + 1.2886 \times PM_{2.5} \quad (10)$$

$$O_3 = 95.039 - 0.22434 \times NO_2 + 0.96905 \times PM_{10} - 1.493 \times PM_{2.5} \quad (11)$$

$$NO_2 = 93.9595 + 0.0017985 \times O_3 + 0.15471 \times PM_{10} - 0.17977 \times PM_{2.5} \quad (12)$$

7.3.2 DT models

In this work, the minimum leave size used is 4, and the maximum number of splits is 6. The main reason for using this setting is to simplify the generated tree.

Dataset	DT results	
	RMSE	R^2
NO_2	2.36	0.65
O_3	7.32	0.54
PM_{10}	3.21	0.89
$PM_{2.5}$	3.14	0.85

Table 5.
DT results after imputing missing data.

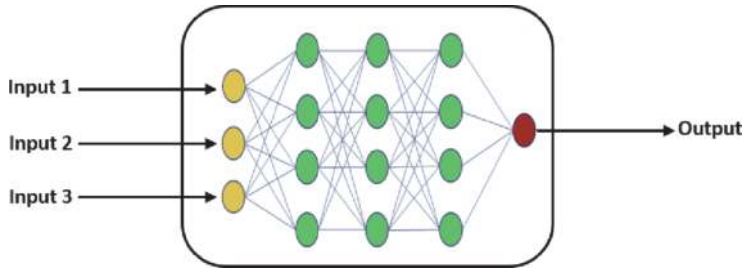

Figure 7.
ANN block diagram structure.

Dataset-	ANN results	
	RMSE	R^2
NO_2	0.05	0.96
O_3	3.25	0.78
PM_{10}	0.33	0.82
$PM_{2.5}$	0.16	0.97

Table 6.
ANN results after imputing missing data.

	Parameters	Value
LRNN	Number of epoch	1000
	Layer delays	1:2
	Hidden sizes	10
	Training function	Back propagation

Table 7.
Parameters setting for LRNN as a regression method.

The obtained models of DT for each dataset were shown in **Figure 6**. **Table 5** explores the obtained results of DT over all datasets.

7.3.3 ANN models

Figure 7 shows the ANN structure used in this chapter, where we have three inputs and a single output. **Table 6** shows the obtained results of

Dataset	LRNN results	
	RMSE	R^2
NO_2	0.22	0.93
O_3	2.76	0.80
PM_{10}	0.02	0.98
$PM_{2.5}$	0.02	0.93

Table 8.
LRNN results after imputing missing data.

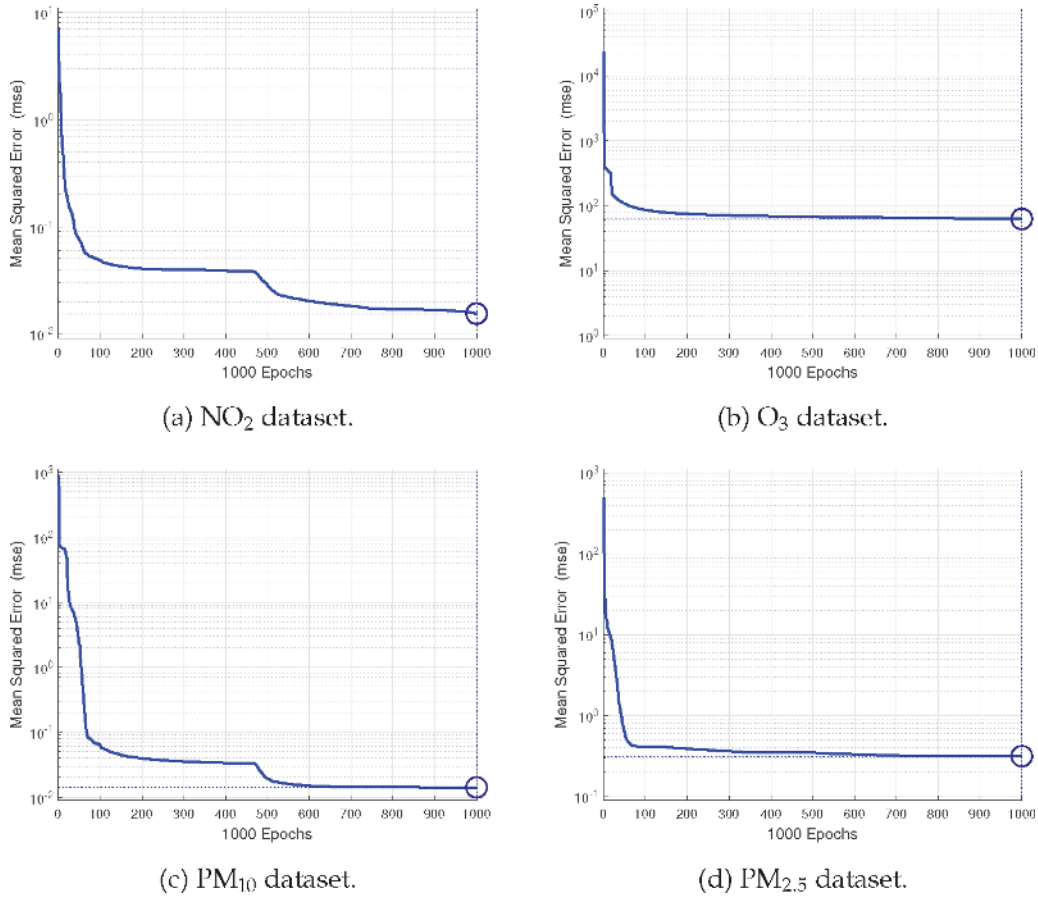

Figure 8.
Convergence curves for LRNN over all datasets.

ANN over all datasets. The performance of ANN is excellent compared to MLR and DT.

7.3.4 LRNN models

In this chapter, we employed the LRNN as a regression model to predict the daily average of air pollutant attributes. **Table 7** shows the parameters setting for LRNN as a regression method. These settings have been selected carefully to fit our data based on a set ore preliminary experiments. **Table 8** shows the obtained results of LRNN. The performance of LRNN is outstanding based on the convergence curves as shown in **Figure 8**. LRNN method can converge within 1000 epochs.

Dataset	Regression model	After imputing		Without imputing	
		RMSE	R^2	RMSE	R2
NO_2	MLR	1.61	0.79	1.79	0.1
	DT	2.36	0.65	1.73	0.16
	LRNN	0.22	0.93	0.26	0.85
	ANN	**0.05**	**0.96**	1.02	0.74
O_3	MLR	1.78	0.82	22.11	0.05
	DT	7.32	0.54	20.44	0.19
	LRNN	**2.76**	**0.80**	10.06	0.61
	ANN	3.25	0.78	8.39	0.76
PM_{10}	MLR	3.84	0.46	3.68	0.83
	DT	3.21	0.89	3.68	0.83
	LRNN	**0.02**	**0.98**	0.3	0.88
	ANN	0.33	0.82	3.85	0.67
$PM_{2.5}$	MLR	1.76	0.62	2.61	0.81
	DT	3.14	0.85	2.57	0.82
	LRNN	**0.02**	**0.93**	2.39	0.9
	ANN	0.16	0.97	2.92	0.85

All significance values are in bold.

Table 9.
Results before and after imputing missing data.

Moreover, the obtained results of LRNN compared to the other previous methods are promising.

7.4 Analysis of the results

Table 9 shows the obtained results before and after imputing missing data. The performance of the LRNN model outperforms other models in three datasets (i.e., NO_2, PM_{10}, and $PM_{2.5}$) based on RMSE and R^2 values. While ANN outperforms other models in O_3 dataset based on RMSE. The performance of ANN over O_3 outperforms other methods. While the performance of MLR is the worst one.

From the obtained results, it can be seen that the performance of the LRNN model has an outstanding performance, where R^2 equals 0.90 in three datasets. However, these obtained results are not perfect since 16.96% of the data is removed for $PM_{2.5}$, and 28.80% of the data is removed from NO_2. Removing the missing data will neglect several records and the dataset may lose important information. **Figure 5** shows the actual and predicted values for NO_2 dataset using all regression methods after imputing missing data.

For more analysis, comparing the obtained results that are reported in **Table 9**, we can notice that the performance of MLR over PM_{10} after imputing the missing data is reduced 19%, while the performance of DT, LRNN, and ANN is improved after imputing missing data for PM_{10} dataset. In general, the performance of the regression models is improved compared to the results reported in **Table 2**.

For example, the R^2 value of ANN over O_3 dataset before imputing missing data was 0.76, and after imputing missing data becomes 0.78, while the RMSE is improved 39%. So, we can conclude that imputing missing data will improve the air quality measurement systems without losing any record of collected data.

8. Conclusion and future work

Data collection from remote sensors suffers from missing data which reduces the overall performance of air quality monitoring systems. Monitoring air pollution is not an easy task, where several measurements are used to evaluate air quality. In this study, four measurements are used to predict air pollution concentrations (i.e., O_3, NO_2, $PM_{2.5}$, and PM_{10}). We imputed the missing data using the Layered recurrent neural network (L-RNN). The performance of four different machine learning models (i.e., LR, DTR, ANN, and L-RNN) was investigated to predict the average daily air pollution concentrations. The performance of the proposed method presented an improvement in the performance of the air quality monitoring system. In future work, we plan to study different methods based on machine learning concepts to enhance the prediction of air pollutant systems. Moreover, we will investigate the general design of the Internet of Things (IoT) applications to improve the performance of the air quality monitoring system.

Acknowledgements

The authors would like to acknowledgement Croatian Meteorological and Hydrological Service for their support.

Author details

Hamza Turabieh[1*], Alaa Sheta[2], Malik Braik[3] and Elvira Kovač-Andrić[4]

1 Department of Information Technology, College of Computers and Information Technology, Taif University, Taif, Saudi Arabia

2 Computer Science Department, Southern Connecticut State University, New Haven, United States of America

3 Department of Computer Science, Al-Balqa Applied University, Salt, Jordan

4 Department of Chemistry, Josip Juraj Strossmayer University of Osijek, Osijek, Croatia

*Address all correspondence to: turabieh@gmail.com

IntechOpen

References

[1] Delfino RJ, Staimer N, Tjoa T, Gillen D, Kleinman MT, Sioutas C, et al. Personal and ambient air pollution exposures and lung function decrements in children with asthma. Environmental Health Perspectives. 2008;**116**(4): 550-558

[2] Belwal C, Sandu A, Constantinescu EM. Adaptive resolution modeling of regional air quality. In: Proceedings of the 2004 ACM Symposium on Applied Computing, SAC '04. New York, NY, USA: ACM; 2004. pp. 235-239

[3] Dastoorpoor M, Goudarzi G, Khanjani N, Idani E, Aghababaeian H, Bahrampour A. Lag time structure of cardiovascular deaths attributed to ambient air pollutants in Ahvaz, Iran, 2008–2015. International Journal of Occupational Medicine and Environmental Health. 2018;**31**(4): 459-473

[4] Adhikari A. Chapter 1 - introduction to spatiotemporal variations of ambient air pollutants and related public health impacts. In: Li L, Zhou X, Tong W, editors. Spatiotemporal Analysis of Air Pollution and Its Application in Public Health. Netherlands: Elsevier; 2020. pp. 1-34

[5] Ghaly A. Mapping environmental pollution, contamination, and waste in the United States. In: Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications. United States: ACM; 2012. p. 41

[6] Chen Y, Wild O, Conibear L, Ran L, He J, Wang L, et al. Local characteristics of and exposure to fine particulate matter (pm2.5) in four Indian megacities. Atmospheric Environment: X. 2020;**5**:100052

[7] Gualtieri M, Øvrevik J, Holme JA, Perrone MG, Bolzacchini E, Schwarze PE, et al. Differences in cytotoxicity versus pro-inflammatory potency of different pm fractions in human epithelial lung cells. Toxicology In Vitro. 2010;**24**(1):29-39

[8] Milojevic A, Wilkinson P, Armstrong B, Bhaskaran K, Smeeth L, Hajat S. Short-term effects of air pollution on a range of cardiovascular events in England and wales: Case-crossover analysis of the minap database, hospital admissions and mortality. Heart. 2014; **100**(14):1093-1098

[9] Dastoorpoor M, Sekhavatpour Z, Masoumi K, Mohammadi MJ, Aghababaeian H, Khanjani N, et al. Air pollution and hospital admissions for cardiovascular diseases in Ahvaz, Iran. Science of the Total Environment. 2019; **652**:1318-1330

[10] Noel De Nevers. Air Pollution Control Engineering. Waveland Press. 2010

[11] Nowak DJ, Hirabayashi S, Doyle M, McGovern M, Pasher J. Air pollution removal by urban forests in canada and its effect on air quality and human health. Urban Forestry & Urban Greening. 2018;**29**:40-48. Wild urban ecosystems: challenges and opportunities for urban development

[12] Kovać-Andrić E, Sheta A, Faris H, Gajdosik MS. Forecasting ozone concentrations in the east of Croatia using nonparametric neural network models. Journal of Earth System Science. 2016;**125**(07)

[13] Sarwar G, Godowitch J, Henderson BH, Fahey K, Pouliot G, Hutzell WT, et al. A comparison of atmospheric composition using the carbon bond and regional atmospheric chemistry mechanisms. Atmospheric Chemistry and Physics. 2013;**13**(19):9695-9712

[14] Sheta A, Faris H, Rodan A, Kovač-Andrić E, Al-Zoubi A. Cycle reservoir with

regular jumps for forecasting ozone concentrations: Two real cases from the east of Croatia. Air Quality, Atmosphere and Health. 2018;**11**(03):559-569

[15] Fuks KB, Woodby B, Valacchi G. Skin damage by tropospheric ozone. Der Hautarzt. 2019:1-5

[16] Lange SS, Mulholland SE, Honeycutt ME. What are the net benefits of reducing the ozone standard to 65 ppb? An alternative analysis. International Journal of Environmental Research and Public Health. 2018;**15**(8)

[17] Isiugo K, Jandarov R, Cox J, Ryan P, Newman N, Grinshpun SA, et al. Indoor particulate matter and lung function in children. Science of the Total Environment. 2019;**663**:408-417

[18] Faustini A, Stafoggia M, Williams M, Davoli M, Forastiere F. The effect of short-term exposure to o3, no2, and their combined oxidative potential on mortality in Rome. Air Quality, Atmosphere and Health. 2019;**12**(5):561-571

[19] Kim C, Hu S-C. Total respiratory tract deposition of fine micrometer-sized particles in healthy adults: Empirical equations for sex and breathing pattern. Journal of Applied Physiology. 2006;**101**:401-412

[20] Deng Q, Lu C, Li Y, Sundell J, Norbäck D. Exposure to outdoor air pollution during trimesters of pregnancy and childhood asthma, allergic rhinitis, and eczema. Environmental Research. 2016;**150**:119-127

[21] Ul-Saufie A, Yahya A, Ramli N, Hamid H. Robust regression models for predicting PM10 concentration in an industrial area. International Journal of Engineering and Technology. 2012;**2**(3): 364-370

[22] Holgate ST, Koren HS, Samet JM, Maynard RL. Air Pollution and Health. United States: Elsevier; 1999

[23] Pokric B, Kreo S, Drajic D, Pokric M, Jokic I, Stojanovic MJ. Ekonet - environmental monitoring using low-cost sensors for detecting gases, particulate matter, and meteorological parameters. In: 2014 Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing. United Kingdom: IMIS-2014, Conference Publishing Service (CPS); 2014. pp. 421-426

[24] Wang F, Liu J. Networked wireless sensor data collection: Issues, challenges, and approaches. IEEE Communication Surveys and Tutorials. 2011;**13**(4):673-687

[25] Turabieh H, Abu Salem A, Abu-El-Rub N. Dynamic L-RNN recovery of missing data in iomt applications. Future Generation Computer Systems. 2018;**89**:575-583

[26] Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architectures. Neural Computation. 2019;**31**(7):1235-1270

[27] Choi E, Schuetz A, Stewart W, Sun J. Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association. 2016; **24**:ocw112

[28] Oeda S, Kurimoto I, Ichimura T. Time series data classification using recurrent neural network with ensemble learning. In: Gabrys B, Howlett RJ, Jain LC, editors. Knowledge-Based Intelligent Information and Engineering Systems. Berlin Heidelberg: Springer; 2006

[29] Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Scientific Reports. 2016;**8**:06

[30] Momeni A, Pincus M, Libien J. Imputation and Missing Data. United

States: Springer International Publishing; 2018. pp. 185-200

[31] Lang KM, Little TD. Principled missing data treatments. Prevention Science. 2018;**19**(3):284-294

[32] Mary IPS, Arockiam L. Imputing the missing data in iot based on the spatial and temporal correlation. In: 2017 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC). Netherlands: Elsevier; 2017. pp. 1-4

[33] Sta HB. Quality and the efficiency of data in "smart-cities". Future Generation Computer Systems. 2017;**74**: 409-416

[34] Feng X, Wu S, Liu Y. Imputing missing values for mixed numeric and categorical attributes based on incomplete data hierarchical clustering. In: Xiong H, Lee WB, editors. Knowledge Science, Engineering and Management. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. pp. 414-424

[35] Sen S, Das M, Chatterjee R. Estimation of incomplete data in mixed dataset. In: Sa PK, Sahoo MN, Murugappan M, Wu Y, Majhi B, editors. Progress in Intelligent Computing Techniques: Theory, Practice, and Applications. Singapore: Springer Singapore; 2018. pp. 483-492

[36] Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. IEEE Access. 2017;**5**: 8869-8879

[37] Perez P, Gramsch E. Forecasting hourly pm2.5 in santiago de chile with emphasis on night episodes. Atmospheric Environment. 2016;**124**: 22-27

[38] Laña I, Del Ser J, Padró A, Vélez M, Casanova-Mateo C. The role of local urban traffic and meteorological conditions in air pollution: A data-based case study in Madrid, Spain. Atmospheric Environment. 2016;**145**: 424-438

[39] Kamińska JA. The use of random forests in modeling short-term air pollution effects based on traffic and meteorological conditions: A case study in wrocław. Journal of Environmental Management. 2018;**217**:164-174

[40] Kamińska JA. Probabilistic forecasting of nitrogen dioxide concentrations at an urban road intersection. Sustainability. 2018;**10**: 4213

[41] Shang Z, Deng T, He J, Duan X. A novel model for hourly pm2.5 concentration prediction based on cart and eelm. Science of the Total Environment. 2019;**651**:3043-3052

[42] Braik M, Sheta A, Al-Hiary H. Hybrid neural network models for forecasting ozone and particulate matter concentrations in the Republic of China. 13. Air, Quality, Atmosphere, and Health. 2020;**13**:839-851. Springer

[43] Sheta AF, Ghatasheh N, Faris H. 2015 6th International Conference on Information and Communication Systems (ICICS). Forecasting global carbon dioxide emission using auto-regressive with eXogenous input and evolutionary product unit neural network models. 2015;182-187. DOI: 10.1109/IACS.2015.7103224

[44] Dotse S-Q, Petra MI, Dagar L, De Silva LC. Application of computational intelligence techniques to forecast daily pm10 exceedances in Brunei Darussalam. Atmospheric Pollution Research. 2018;**9**(2):358-368

[45] Sun W, Sun J. Daily pm2.5 concentration prediction based on principal component analysis and lssvm optimized by cuckoo search algorithm.

Journal of Environmental Management. 2017;**188**:144-152

[46] Xu Y, Du P, Wang J. Research and application of a hybrid model based on dynamic fuzzy synthetic evaluation for establishing air quality forecasting and early warning system: A case study in China. Environmental Pollution. 2017; **223**:435-448

[47] Luo H, Wang D, Yue C, Liu Y, Guo H. Research and application of a novel hybrid decomposition-ensemble learning paradigm with error correction for daily pm10 forecasting. Atmospheric Research. 2018;**201**:34-45

[48] Aznarte JL. Probabilistic forecasting for extreme no2 pollution episodes. Environmental Pollution. 2017;**229**: 321-328

[49] Wang D, Wei S, Luo H, Yue C, Grunder O. A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. Science of the Total Environment. 2017; **580**:719-733

[50] Kumar A, Goyal P. Forecasting of air quality in Delhi using principal component regression technique. Atmospheric Pollution Research. 2011;**2** (4):436-444

[51] Akhtar A, Masood S, Gupta C, Masood A. Prediction and analysis of pollution levels in Delhi using multilayer perceptron. In: Satapathy SC, Bhateja V, Raju KS, Janakiramaiah B, editors. Data Engineering and Intelligent Computing. Singapore: Springer Singapore; 2018. pp. 563-572

[52] Yadav V, Nath S. Identification of relevant stochastic input variables for prediction of daily pm10 using artificial neural networks. In: Ray K, Sharma TK, Rawat S, Saini RK, Bandyopadhyay A, editors. Soft Computing: Theories and Applications. Singapore: Springer Singapore; 2019. pp. 23-31

[53] Singh P. Linear Regression. Berkeley, CA: Apress; 2019. pp. 43-64

[54] Wang S, Huang GH, He L. Development of a clusterwise-linear-regression-based forecasting system for characterizing dnapl dissolution behaviors in porous media. Science of the Total Environment. 2012;**433**:141-150

[55] Swetapadma A, Yadav A. A novel decision tree regression-based fault distance estimation scheme for transmission lines. IEEE Transactions on Power Delivery. 2017;**32**(1):234-245

[56] Qin H, Gong R, Liu X, Bai X, Song J, Sebe N. Binary neural networks: A survey. Pattern Recognition. 2020;**105**:107281