Computer Science and Information Systems 23(3):1077-1102 doi: https://doi.org/10.2298/CSIS251101031K

# Artificial Neural Network Modeling for Air Pollution Prediction: LSTM versus the Levenberg-Marquardt Approach

Goran Keković[1], Rade Božović[1], Sonja Ketin[2], Vladimir Mikić[1], Miloš Ilić[1] and Boban Vesin[3]

[1] Faculty of Information Technology, Alfa BK University, 11000 Belgrade, Serbia
goran.kekovic@alfa.edu.rs (corresponding author),
{rade.bozovic, vladimir.mikic, milos.ilic}@alfa.edu.rs
[2] School of Railroad Transport, Academy of Technical and Art Applied Studies, 11000 Belgrade Serbia
sonja.ketin@vzs.edu.rs
[3] School of Business, University of South-Eastern Norway, Raveien 215, Borre, Vestfold, 3184, Norway
boban.vesin@usn.no

**Abstract.** Accurate prediction of air pollutant concentrations remains a critical challenge for environmental monitoring and public health, demanding robust and adaptive artificial intelligence approaches. This study investigates the effectiveness of various types of artificial neural networks (ANNs), including Long Short-Term Memory networks (LSTM) and networks based on the Levenberg-Marquardt algorithm (LM) and its variant with Bayesian regularization (LMBR), in predicting air pollution under different data conditions. Since LSTM networks are based on first derivative loss function algorithms and the LM algorithm is usually superior to this type of algorithm, a comparison of these networks was conducted. This is further supported by the limited coverage of this topic in the existing literature. ANNs were tested on two different datasets: the Air Quality dataset, where the target variable was the concentration of benzene ($C_6H_6$) and the Beijing PM2.5 dataset, where the target was the concentration of PM2.5 particles. The performance metrics of the ANNs were the root mean square error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE). It is shown that, in the case of the Air Quality dataset, the values of these parameters $RMSE = 0.11$ $\mu$g/m$^3$, $MAE = 0.09$ $\mu$g/m$^3$, $MAPE = 1\%$ for LSTM networks and $RMSE = 0.14$ $\mu$g/m$^3$, $MAE = 0.092$ $\mu$g/m$^3$, $MAPE = 1\%$ for LMBR networks, were competitive. For LM networks, these values were significantly higher: $RMSE = 0.57$ $\mu$g/m$^3$, $MAE = 0.2$ $\mu$g/m$^3$, $MAPE = 2\%$. Contrastingly, in the case of the Beijing database, the values of all parameters were drastically higher: $RMSE = \{45.74$ $\mu$g/m$^3$, $64.94$ $\mu$g/m$^3$, $65.68$ $\mu$g/m$^3\}$, $MAE = \{30.32$ $\mu$g/m$^3$, $42.83$ $\mu$g/m$^3$, $54.5$ $\mu$g/m$^3\}$ and $MAPE = \{52\%, 72\%, 74.25\%\}$ for LSTM, LMBR, and LM networks, respectively. In this case, the benzene concentration values exhibited a strictly linear correlation with the input variables. For the Beijing dataset, the relationships between PM2.5 concentration and predictor values were non-monotonic, leading to a drastic drop in the performance of all networks. Findings reveal that, in this case, LSTM networks were more robust compared to LMBR and LM networks, as the values of their RMSE, MAE, and MAPE parameters were

significantly lower. Furthermore, it is shown that the input variable selection technique (IVS) can be used to detect seasonal trends in the input data.

**Keywords:** Artificial Intelligence, LSTM Network, Levenberg-Marquardt Algorithm, Input Variable Selection.

## 1. Introduction

Air pollution is becoming one of the major problems facing modern human society due to various factors, such as the combustion of fossil fuels in industry, vehicle exhaust emissions associated with transport, and the migration of the population to urban areas [35, 39]. In general, all components of air pollution can be divided into two groups: primary and secondary. Primary components are created through the direct combustion of fossil fuels, resulting in the release of gases and particulate matter into the atmosphere. The most common representatives of this group are particulate matter (PM10, PM2.5), nitrogen and sulfur oxides ($NO_x$, $SO_x$), non-methane hydrocarbons (NMHC) and ozone $O_3$ [38]. Secondary components are formed by chemical reactions between primary components. A typical example of this group is the conversion of non-methane hydrocarbons into tropospheric ozone and peroxyacetyl nitrate (PAN) in the presence of the photocatalytic action of solar radiation. Numerous studies have examined the harmful effects of pollutants on human health, with PM being of particular importance [24, 28, 70, 54].

Although PM air pollutants have been addressed in a relatively large number of papers, benzene has received limited attention. This is unjustified, considering that exposure to benzene can cause various types of cancer in humans [47]. The adverse effects of this and other pollutants have led to the development of modern technologies to monitor their concentrations in the air, with artificial intelligence (AI) methods playing a particularly important role, because they are not subject to the shortcomings typically associated with traditional regression methods with regard to input data structure (the problem of multicollinearity and types of variables, mutual independence of data, etc.) [22]. In general, AI methods used for air pollution prediction can be divided into five categories: fuzzy logic [8], hidden Markov models [30], ensemble models [13], artificial neural networks (ANNs) [5], and deep learning [24]. In the context of this paper, the last two categories are of greatest interest.

ANN algorithms are widely used in air quality control, and recently, LSTM deep neural networks have gained particular popularity [74, 29, 49]. The popularity of these networks is due to their ability to predict short-term and long-term dependencies of the targeted variable, i.e. pollutant concentration. Various types of these networks have been applied, and one study found that Bidirectional Long Short-Term Memory (Bi-LSTM) achieves the best results [71]. In another interesting study, the authors developed a model to predict PM particle concentrations ten days in advance in Seoul, South Korea, using two techniques: LSTM and deep autoencoder [64]. In this paper, it was shown that LSTM networks outperformed DAE, further demonstrating the advantage of LSTM. Recently, researchers have increasingly used LSTM deep learning algorithms, which have proven to be very effective in processing big data and time series [65, 53, 1, 27]. The disadvantage of these algorithms is the need to adjust the parameters in their hyperspace, which is time-consuming, and for this purpose, particle swarm optimization and genetic algorithms are used [68].

Although recurrent artificial neural networks (RNNs) represent a modern type of ANNs for the prediction of air pollution, the LM algorithm and its Bayesian-regularized variant LMBR were also considered in this study. The reason is that, for small and medium datasets, algorithms based on the second derivative of the loss function are generally more efficient than those based on the first derivative [56]. Accordingly, the competitiveness of the LM and LMBR algorithms can also be expected, since all RNN variants rely not only on memory cells, but also on algorithms based on the first derivative of the loss function. LSTM and other variants of RNNs can learn temporal correlations between input variables due to their property of memorizing long-term effects. On the other hand, algorithms based on the second derivative work in batch or semi-batch mode, which means that the history of the time series values of the predictors is taken into account. Another reason why their competitiveness is worth examining is that there are almost no studies comparing the properties of LSTM and LM/LMBR networks in air pollution tasks. A relevant study has explored this topic in a different context. Specifically, the authors combined LSTM networks with the LM algorithm to identify unknown aerodynamic parameters using real flight data from aircraft [73]. Their results demonstrated the effectiveness of this approach compared to the parameter values obtained from wind tunnel experiments.

In the mentioned studies, the term air pollution prediction often refers to the prediction of the concentration of specific air pollutants. The forecast horizon usually varies from one to several days, and the discrepancies between the simulated and actual values of the concentration of a given pollutant are reported most frequently [36, 41]. The output variable (pollutant concentration) is a function of numerous input parameters, so the question that arises is the accuracy of this approach. For example, in all AI-based simulations, meteorological conditions are regularly taken into account as part of the input parameters, which are highly variable both locally and globally. Therefore, the horizon of future values is usually determined on the basis of the average values of the parameters in the past. When using ANNs, the predicted concentrations of pollutants could be grouped into specific intervals or classes, since ANNs are more efficient at classification tasks. However, this approach can estimate the interval of expected values, which can have a significant impact on the planning of daily activities.

At the same time, the objective was to compare the properties of these models to identify the most suitable approach for the prediction of air pollution. However, like other AI-based methods, they remain sensitive to the structure of the input data, and this issue therefore requires particular attention. This limitation can be partially addressed through input variable selection (IVS). According to recent studies, IVS techniques can be classified into three categories [7].

In the first category, known as filter techniques, the input data are preprocessed using statistical analysis methods and algorithms that determine the relevance of the input variables [18]. A typical representative of this category is the algorithm mrMR (mr-minimum redundancy, MR-maximum relevance) based on the application of the concept of mutual information [42]. This category is the simplest to implement and is independent of the AI method to which it passes the selected variables. IVS methods that involve an extensive search through the parameter space of the input data, monitoring the effect of adding and removing input variables, are called wrapper methods [25]. In this sense, there are two types of these methods: sequential addition of variables to the initial empty set (SFS) and sequential elimination of variables from the entire initial set of variables (SBS). In both

cases, the effects of adding or removing variables are monitored by the behavior of the loss function. This category of methods is technically much more demanding than the first category in terms of computing time and memory resources.

The final category comprises the most complex methods in terms of practical implementation, as they are applied during the training phase of the AI method to which the variables are assigned. This group of methods is called embedded methods and it should be noted that the situation is particularly complex when it comes to ANNs [51]. Typical representatives of this group are the CART and ID3 algorithms in decision trees. The application of IVS yields benefits in terms of reduced computation time and lower memory requirements, but it does not completely resolve the fundamental problem of the dependence of AI methods on the structure of the input data. Among these factors, the sample size is one of the most influential. The widely accepted view is that a larger sample leads to greater accuracy of these methods, due to a larger training sample. However, this view cannot always be accepted as correct, because there are studies in which researchers have shown that high accuracy of AI methods can be achieved even on small sample sizes [52]. Moreover, a further increase in the sample size may even reduce the accuracy of AI methods [57].

Despite the growing use of AI methods in air pollution monitoring and prediction, several important issues remain insufficiently explored. In particular, the comparative performance of different neural network architectures under limited data conditions, as well as the role of advanced preprocessing techniques in enhancing model performance, have yet to be fully clarified. These challenges are especially relevant in real-time monitoring scenarios, where accurate modeling of the relationship between environmental variables and pollutant concentrations is essential.

Motivated by these considerations, this study addresses the following research questions:

- **RQ1**: How competitive are LSTM and LM/LMBR neural network models in real-time air pollution monitoring, particularly in modeling the relationship between the current values of the target variable and the corresponding input variables?
- **RQ2**: To what extent can the IVS method be effectively applied in the data preprocessing stage for air pollution monitoring tasks, including the identification of relevant input features and seasonal patterns in the data?

Given the high variability of input parameters in air pollution monitoring—such as meteorological conditions, geographical characteristics, and temporal dynamics—as well as the wide range of available AI-based modeling approaches, deriving universally valid conclusions remains challenging. Nevertheless, the results obtained in this study provide valuable insights into the applicability of the considered methods and can serve as a guidelines for future research in this domain.

The main contributions of this study can be summarized as follows:

- A comparative evaluation of LSTM and LM/LMBR networks is conducted using the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) metrics, demonstrating that their predictive performance is largely comparable when trained on relatively small datasets.
- The study shows that monotonicity relations in input data lead to a significant improvement in the predictive performance of the considered neural network models.

- The robustness of LSTM networks to nonlinear relationships between variables has been generalized and extended to non-monotonic relationships.
- The applicability of the IVS technique is further extended by demonstrating its effectiveness in identifying and analyzing seasonal patterns present in the input data.

The remainder of this paper is organized as follows: Section 2 reviews the related work and provides the broader context for the study. Section 3 describes the methods used in the research, including the research approach and procedures. Section 4 presents and discusses the results obtained from the analysis. Section 5 outlines the main limitations of the conducted research. Finally, Section 6 concludes the paper by summarizing the key findings and suggesting directions for future work.

## 2. Related work

Air pollution prediction has emerged as a highly active research domain within environmental informatics, given the paramount importance of accurate forecasts in safeguarding public health and informing policy decisions. A substantial body of research has demonstrated the potential of deep learning models, particularly LSTM networks, in capturing long-range temporal dependencies in nonlinear and noisy time series data. Numerous studies have utilized LSTM architectures to forecast particulate matter concentrations, ozone, and other pollutants, consistently demonstrating superior performance compared to traditional regression and shallow learning models. For example, in [44], the author demonstrated that LSTM models outperformed classical regression approaches for PM2.5 forecasting. Similarly, the authors in [12, 61] introduced spatiotemporal variants such as convolutional LSTM and recursive LSTM to integrate both temporal dynamics and spatial correlations for air pollution prediction, demonstrating significant reductions in prediction errors. Further improvements have been achieved by hybrid deep learning methods. The authors in [31] proposed a CNN-LSTM framework for PM2.5 prediction, while the authors in [69] employed an attention-based LSTM for real-time pollution monitoring, with both studies reporting lower RMSE and MAPE values compared to baseline models. Complementing these views, the authors in [58] also demonstrated that LSTM networks successfully internalize complex, non-linear temporal dependencies of ambient quality metrics. These findings have reinforced the prevailing view that LSTM architectures are the state of the art in air pollution prediction tasks.

Despite this dominance, evidence suggests that optimization methods based on the LM algorithm and its Bayesian regularization variant, LMBR, remain competitive in certain contexts, especially when there are strong linear correlations between predictors and target variables. In their study [15], the authors demonstrated that LM-ANNs outperformed both multiple linear regression and other training algorithms in PM2.5 prediction in India, achieving the highest coefficient of determination ($R^2 = 0.8164$) and the lowest RMSE (9.52). Similar results were obtained in energy demand forecasting, where LM-ANN models achieved superior accuracy compared to scaled conjugate gradient and regression methods [62, 56]. Research reported in [26] analyzed multiple time-series neural network models, including various configurations, suggesting that the Levenberg–Marquardt approach can be effective for Air Quality Index forecasting based on the year of various gas measurements. In [4], the authors conducted a comparative analysis of ANN models for long-term forecasting of PM10 concentrations. Their findings confirmed

that, despite the low RMSE values achieved by LM-ANN, the LMBR approach consistently delivered the most reliable overall performance, thus underscoring its potential in air quality forecasting applications. The presented studies indicate that LM and LMBR algorithms can be highly effective in domains where predictors exhibit strong linear relationships with the response variable and where robustness to smaller sample sizes is required. Direct comparative studies of the LSTM and LM/LMBR methods in the domain of air pollution remain scarce, but emerging results point to their potential complementarity and indicate that such comparisons may offer valuable insights for identifying the most effective predictive approaches, thus contributing to improved public health through more accurate air quality management.

In summary, while LSTM networks have established themselves as the dominant method for air pollution prediction due to their ability to handle complex nonlinear and temporal dependencies, LM and LMBR-based networks remain important alternatives that can offer competitive accuracy in contexts characterized by stronger linear dependencies. The comparison of these approaches represents a promising research direction that ensures both predictive accuracy and interpretability. This need forms the foundation of the presented study, which systematically evaluates the comparative performance of LSTM and LMBR in real-world air quality monitoring tasks.

## 3. Methods

To facilitate a clearer understanding of the methods and materials used in this study, an overview pseudocode of the entire research process is presented below. This pseudocode is intended to summarize the main methodological steps of the proposed air pollution modeling framework and to provide a structured guide for interpreting the individual phases of model development, validation, and evaluation. Each step is subsequently described in greater detail in the following subsections. In addition to improving the readability of the methodology, this overview also highlights the sequential dependencies between data preparation, model construction, validation, fine-tuning, and final performance assessment.

The presented pseudocode shows that the overall process of air pollution modeling using ANNs can be divided into four main phases. While Phases 1, 2, and 4 are relatively straightforward, Phase 3 requires additional clarification due to its greater methodological complexity. In Phase 3, for each of the considered network models - LSTM and LMBR/LM - the most appropriate partitioning of the input dataset into k folds is determined using k-fold cross-validation. This procedure is a standard technique for mitigating overfitting during model training. Once the optimal fold scheme has been established, the resulting models are further refined through a fine-tuning process. This step involves adjusting both the network architecture and the associated hyperparameters (lines 9–12 of the pseudocode) in order to improve model performance. Finally, lines 13–16 of the pseudocode correspond to the prediction and evaluation stage, during which the trained models generate output predictions and their predictive performance is assessed. Each phase of the pseudocode will be described in greater detail in the sections that follow.

**Algorithm 1**. ANN Modeling and Prediction Framework

**Input:** Raw dataset $D$; initial ANN architectures $\{LSTM, LMBR/LM\}$
**Output:** Predictions $\bar{Y}$ and performance metrics $M$ for each ANN model

**Phase 1: Data Preprocessing**
Remove incomplete observations from $D$
Normalize variables using z-score standardization
$x_{\text{norm}} = (x - \mu)/\sigma$
where $\mu$ is the mean value and $\sigma$ is the standard deviation
Apply noise filtering to obtain cleaned dataset $D\_clean$

**Phase 2: Feature Extraction**
Apply IVS technique to $D\_clean$ to identify seasonal trends
Construct feature set $F$

**Phase 3: Model Training and Optimization**
**for each** $ANN\_model \in \{LSTM, LMBR/LM\}$ **do**
    Determine optimal $k$ using $k$-fold cross-validation
    **repeat**
        Tune hyperparameters of $ANN\_model$
        Adjust network architecture
        Evaluate model using $k$-fold cross-validation
    **until** convergence of performance metric

**Phase 4: Prediction and Evaluation**
    Generate predictions $\bar{Y}$ using trained $ANN\_model$
    Compute performance metrics $M$
**end for**
**return** $\bar{Y}$ and $M$

### 3.1. Databases and input variable selection

As already noted, there is no definitive classification of AI methods with respect to the structure of input data; most recommendations rely on empirical findings from numerical experiments. A common starting point is the analysis of linear correlations between the target variable and the predictors. Since linear relationships represent a special case of monotonic dependence, their presence or absence often indicates whether the underlying relationships are predominantly monotonic or whether they are more complex and nonlinear.

The two databases used in this study, namely the Air Quality dataset and the Beijing dataset, were intentionally selected because they represent contrasting structures of variable relationships. The Air Quality dataset is dominated by monotonic relationships between variables, whereas the Beijing dataset contains predominantly non-monotonic relationships among input predictors. Examining these opposing structural characteristics enables a more informative assessment of research questions RQ1 and RQ2.

However, AI methods alone are not sufficient for a comprehensive characterization of these relationships; therefore, IVS techniques are also applied as part of the raw data preprocessing stage.

**Air Quality dataset** The numerical experiments were conducted using a dataset collected from air pollution measurements in an Italian city between March 2004 and February 2005 [59]. The database contained 9358 samples, with values averaged over hourly intervals, obtained from the sensor system. The target variable was the concentration of benzene $C_6H_6$. Some of the recorded values were invalid (e.g. negative concentrations, negative temperature, and humidity), most likely due to sensor drift and other measurement-related issues. An initial approach was to remove all samples that contained such invalid values; however, this would have reduced the input dataset to only 827 samples. Because this reduction was considered unacceptable, the distribution of invalid data was then examined for each variable. Based on this analysis, the variables were divided into three groups, with 3.91%, 17.99% and 90% invalid data, respectively. Variables with more than 3.91% invalid data were excluded from further analysis. As a result, the number of input variables, or predictors, was reduced from 12 to 8. It is important to note that removing these variables did not affect the research questions introduced earlier, since the corresponding measurements were available from other locations and could therefore still be represented in the input dataset. For the remaining variables, invalid values were replaced with the mean of all valid observations in the time series. The statistical parameters of the filtered data are given in Table 1. As shown in the table, some variables are expressed in arbitrary units, which corresponds to sensors nominally intended to measure the concentrations of specific air pollutants (PT08S1, PT08S2, etc.).

In addition, all variables were found to be normally distributed. Therefore, Pearson's correlation test was suitable for correlation analysis. Since the target variable was the concentration of benzene, Table 2 presents the statistically significant values of the correlation coefficient between benzene and other variables, at a statistical significance level of 5%. A visual inspection of Table 2 shows that the values of the correlation coefficient are very high. The correlations between the input variables were also high, indicating the presence of multicollinearity, which reflects the synergistic effect of various air pollutants in their mixture in the air. Since the correlation values between the input variables are high, it can be logically assumed that this parameter can serve as an indicator of seasonal changes in the predictor-target relationship. Furthermore, it can also be used for input variable selection in the case of incomplete input data from the sensor network.

**Table 1.** Statistical parameters of collected data

| Parameters | Unit | Min. | Max. | Std. |
|---|---|---|---|---|
| PT08.S1(CO) | – | 647.25 | 2040 | 212.80 |
| $C_6H_6$(GT) | $\mu$g/m$^3$ | 0.15 | 63.74 | 7.30 |
| PT08.S2(NMHC) | – | 383.25 | 2214 | 261.56 |
| PT08.S3(NOx) | – | 322 | 2683 | 251.74 |
| PT08.S4(NO2) | – | 551 | 2775 | 339.36 |
| PT08.S5(O3) | – | 221 | 2522.8 | 390.61 |
| T | $^\circ$C | 0.05 | 44.6 | 8.63 |
| RH | % | 9.18 | 88.72 | 16.97 |
| AH | g/m$^3$ | 0.18 | 2.23 | 0.40 |

**Table 2.** Correlation coefficients with $C_6H_6$

| Parameters | Correlation with $\mathbf{C_6H_6}$ |
|---|---|
| PT08.S1(CO) | 0.85 |
| PT08.S2(NMHC) | 0.77 |
| PT08.S3(NOx) | 0.51 |
| PT08.S4(NO2) | 0.77 |
| PT08.S5(O3) | 0.64 |
| T | 0.97 |
| RH | 0.93 |
| AH | 0.98 |

The mrMR algorithm was employed [43]. This algorithm selects input variables that are minimally correlated (minimal redundancy, mr) with each other and maximally correlated (maximal relevance, MR) with benzene as the output variable. The mrMR algorithm relies on mutual information, defined by the following formula:

$$I = \sum_{x,y} p(x,y) \cdot \log \frac{p(x,y)}{p(x) \cdot p(y)} \tag{1}$$

where $p(x,y)$ is the joint distribution of the variables $x$ and $y$, while $p(x)$ and $p(y)$ are the marginal distributions of the respective variables. The meaning of mutual information can be seen if $X \cap Y = \{0\}$ is considered, where $p(x,y) = p(x) \cdot p(y)$, and from the above formula it follows that $I(X,Y) = 0$. Mutual information represents the amount of information about one variable that is related to another variable. In practical applications, the mrMR algorithm tries to maximize the $MI$ score:

$$MI = \frac{1}{S} \cdot \frac{\sum_{x \in S} I(x,y)}{\sum_{x,z \in S} I(x,z)} \tag{2}$$

In the above equation, the numerator represents the sum of the mutual information of the predictor $x$ with the target variable $y$, and the denominator represents the sum of the mutual information between the predictors. The mrMR algorithm attempts to maximize the score by finding a suitable subset $S$ of input variables, so that the numerator in the above formula is maximal (relevance) and the denominator is minimal (redundancy).

The results shown in Fig. 1 indicate seasonal trends for the winter and summer seasons. In the summer season, NMHC, temperature, and absolute humidity are identified as the dominant variables. This suggests enhanced benzene formation through secondary chemical reactions among primary pollutants under the photocatalytic action of solar radiation. During the winter season (Fig. 1D), nitrogen oxide emerges as the dominant variable, most likely due to increased fossil-fuel combustion for heating, while temperature and air humidity also appear among the relevant variables. In the other seasons (spring and autumn, Fig. 1A, 1B), no significant predictors were identified. These findings are further supported by Fig. 2, which shows two distinct peaks in the daily benzene concentration profile during spring and summer. These peaks occur in the morning and late afternoon,

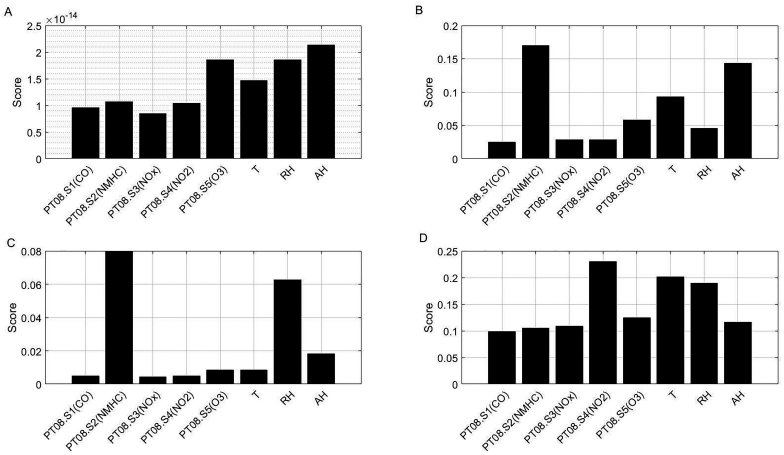


**Fig. 1.** Seasonal trends in the influence of the predictor on the concentration of benzene: A-Spring, B-summer, C-autumn, D-winter

most likely due to increased traffic during the morning and afternoon rush hours. At the same time, they are superimposed on increased benzene concentrations caused by the previously mentioned secondary reactions in the summer season (Fig. 2B) and by thermal combustion during the winter season (Fig. 2D).

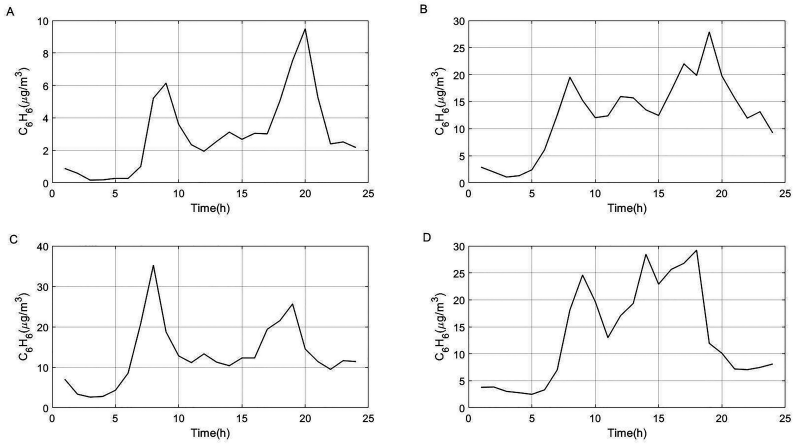


**Fig. 2.** Typical daily trends in benzene: A-Spring, B-summer, C-autumn, D-winter

**Beijing PM2.5 dataset** This database comprised 43825 measurements of PM2.5 concentrations recorded at one-hour intervals in Beijing, China, together with the corresponding meteorological conditions [10]. Data were recorded between 2010 and 2015. After removing incomplete or invalid samples (e.g., those affected by sensor drift or containing negative concentrations), the initial dataset was reduced to 41757 samples. The original dataset also included columns containing the month and hour of each recorded sample, which allowed the data to be classified by season even after preprocessing. To avoid the effect of sample size and to ensure comparability with the first database used in this study, the dataset was further reduced to 9358 samples. The target variable was the concentration of PM2.5 particles, while the predictor variables were: Dew Point (DEWP), Temperature (T), Pressure (P), combined wind direction (cbwd), cumulated wind speed (Iws), cumulated snowfall hours (Is) and cumulated rainfall hours (Ir). The main statistical characteristics of these variables are given in Table 3, while Fig. 3 shows the average daily values of the predictors throughout the seasons.

**Table 3.** Statistical parameters of the Beijing dataset

| Parameters | Unit | Min. | Max. | Std. |
|---|---|---|---|---|
| PM2.5 | $\mu$g/m$^3$ | 1 | 980 | 97 |
| DEWP | $^\circ$C | -28 | 28 | 14.91 |
| T | $^\circ$C | -19 | 41 | 12.96 |
| P | hPa | 994 | 1045 | 10.31 |
| cbwd | - | 1 | 4 | 1.28 |
| Iws | m/s | 0.45 | 565.5 | 59.23 |
| Is | - | 0 | 27 | 1.14 |
| Ir | - | 0 | 36 | 1.71 |

In Fig. 3, the dotted line denotes the concentration threshold for PM2.5 that is considered hazardous to human health. According to the China national ambient air quality standards (NAAQS), standard GB 3095-2012, the corresponding permissible average concentration value is 35 $\mu$g/m$^3$ (Grade I) [72]. As shown in Fig. 3, the average daily concentration of PM2.5 exceeds this threshold in all seasons. This exceedance is most pronounced in the summer (Fig. 3C) and autumn seasons (Fig. 3D). The elevated concentrations of PM2.5 observed during summer are most likely associated with high temperatures and the photocatalytic effect of solar radiation, which promote reactions among gaseous pollutants (SO, NOx, VOCs, etc.) and the formation of secondary particles (nitrites, sulfides, etc.). In addition, increased air conditioning use and increased energy demand, together with the evaporation of organic compounds, can contribute further to the condensation and formation of PM2.5 particles. This interpretation is supported by the graph in Fig. 3A, which shows that PM2.5 concentrations are lowest during winter, when both temperature and humidity reach their lowest levels. According to the filtered data, the average winter temperature was 3.8 $^\circ$C. Further preprocessing showed that the linear correlations between the target variable and the predictors were not statistically significant. Kendall's correlation test, performed at the 5% significance level, indicated that the correlation co-

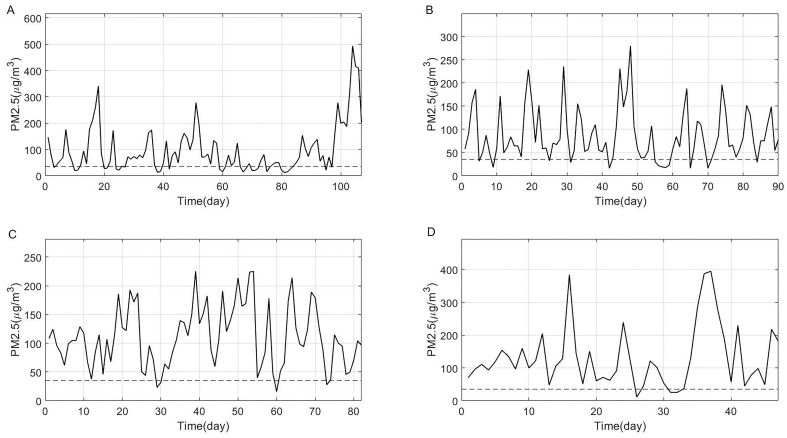


**Fig. 3.** Typical average daily PM2.5 particle concentrations: A-winter, B-spring, C-summer, D-autumn

efficients were negligible or borderline statistically significant, as presented in the table below.

**Table 4.** Correlations of PM2.5 particle concentrations with input variables

| **Parameters** | **Correlation with PM2.5** |
|---|---|
| DEWP | 0.303 |
| TEMP | 0.128 |
| PRESS | -0.252 |
| cwbd | 0.268 |
| Iws | -0.300 |
| Is | -0.007 |
| Ir | 0.010 |

The data listed in the table above lead to the conclusion that the relationships between the variables in the system are nonlinear and complex, which is why it is necessary to use other methods to investigate them. Among the various methods, nonlinear regression is one of the best-known and widely used. Consequently, a nonlinear regression was performed between the target variable $y$ and each predictor $x$ separately, that is:

$$y = a_0 + \sum_{i=1}^{K} a_i x^i \tag{3}$$

where $K$ is the highest degree of the polynomial in Eq. (3). If $K > 1$, it can be assumed that the system is nonlinear, i.e., for $K \gg 1$, the system is highly nonlinear. Fig. 4 shows the results. As can be seen, the degree of nonlinearity $K$ is very high for all predictors, which means that nonlinear relationships prevail between the target variable and the predictors. These results are also consistent with the data in Table 4. In the cases of the DEWP and cwbd predictors, the Kendall coefficient values are borderline statistically significant, since the statistically significant values of the Kendall coefficient $\rho$ occur when $|\rho| > 0.25$.

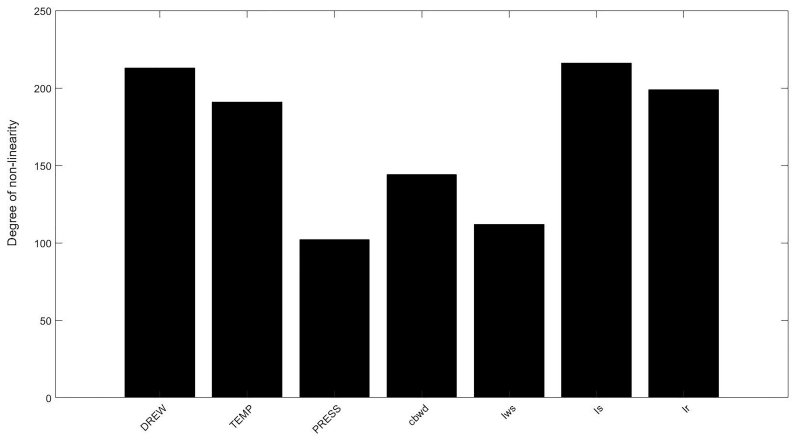


**Fig. 4.** The degree of nonlinearity between the target variable and the predictor

### 3.2. Performance criteria

The efficiency of ANNs in air pollution monitoring tasks is commonly evaluated using the following metrics: RMSE, MAE, and MAPE [3]. These metrics are defined by the following equations:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}} \tag{4}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \tag{5}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{x_i - y_i}{x_i}\right| \cdot 100 \tag{6}$$

### 3.3. Artificial neural networks

For air pollution prediction, LSTM networks and ANNs based on the LM algorithm were employed. LSTM networks, as a class of recurrent neural networks, are widely used in air pollution prediction problems [20]. They consist of conventional neural layers, including one or more characteristic memory layers, or cells. These layers differ in both structure and function, as they store short-term cell states and thereby enable the network to capture long-range dependencies in the data. This makes such networks particularly suitable for tasks that involve sequential data types. The structure of these cells is shown in Fig. 5. Four units with specific functions can be identified. The symbols $\sigma$ and $\tanh$ represent the standard transfer functions sigmoid and hyperbolic tangent, and the operator $\odot$ denotes element-wise matrix multiplication. The operation of the memory layer can be described as follows: a sample vector $X_t$ from the input dataset propagates through the characteristic units $f_t$, $i_t$, $\bar{C}_t$, and $O_t$, also called gates, while interacting with the previous states $C_{t-1}$ and $h_{t-1}$ of the layer (cell), resulting in updated states $C_t$ and $h_t$. The first unit, $f_t$, known as the forget gate, is given by the formula:

$$f_t = \sigma\big(W_f \cdot X_t + R_f \cdot h_{t-1} + b_f\big) \tag{7}$$

where $W_f$, $R_f$, and $b_f$ denote, respectively, the corresponding neural and recurrent weights and the bias matrices. Passing the vector through these gates, i.e., by applying Eq. (7), the previous state of the cell is filtered, i.e., the unnecessary part of the previous state determined by the vector $h_{t-1}$ is discarded or forgotten. Continuing further propagation through the input gate $i_t$, the cell state is prepared to be updated with new information. This is described by the following equation:

$$i_t = \sigma\big(W_i \cdot X_t + R_i \cdot h_{t-1} + b_i\big) \tag{8}$$

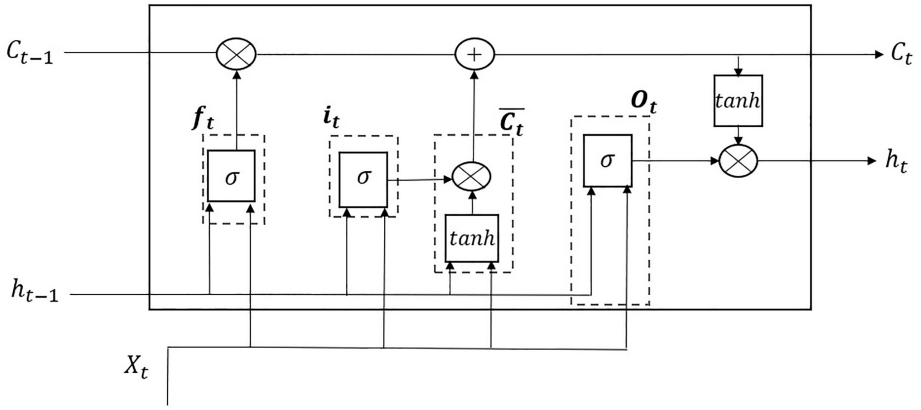


**Fig. 5.** Memory layer in LSTM networks

In the next phase, new information is added to the cell state:

$$C_t = f_t \odot C_{t-1} + i_t \odot \bar{C}_t \tag{9}$$

with a previously prepared candidate $\bar{C}_t$:

$$\bar{C}_t = \tanh\left(W_c \cdot X_t + R_C \cdot h_{t-1} + b_C\right) \tag{10}$$

Applying the output gate $O_t$ then gives:

$$O_t = \sigma\big(W_o \cdot X_t + R_o \cdot h_{t-1} + b_o\big) \tag{11}$$

so that the new cell state vector $h_t$ is finally obtained:

$$h_t = O_t \odot \tanh(C_t) \tag{12}$$

The structure of the LSTM network consisted of three layers: an input layer, an LSTM layer, and an output layer. Because this study considered real-time prediction of the benzene concentration based on the current values of the predictors, a single hidden LSTM layer was sufficient. The input layer, as the initial layer, had a number of neurons equal to the number of predictors, while the output layer had one neuron because the target variable was benzene concentration. The number of neurons in the LSTM layer was determined by optimization, which will be discussed in more detail later in the text.

The LM algorithm is one of the most efficient ANN algorithms for small and medium-sized samples [62, 55, 40]. Its modification with Bayesian regularization makes it particularly suitable for comparison with other algorithms, since this type of regularization introduces a special self-penalization of the model, equivalent to the principle of Occam's razor [34, 16]. This algorithm minimizes the following objective function:

$$F = \beta E_D(D \mid w, M) + \alpha E_w(w \mid M) \tag{13}$$

where $E_D = \frac{1}{M}\sum_{i=1}^{P}(y_i - y_{it})^2$ represents the standard objective function, while $E_w = \frac{1}{N}\sum_{i=1}^{N} w_i^2$ is the sum of the squares of the neural weights, and $\alpha$ and $\beta$ are hyperparameters that need to be determined. The second term in Eq. (13) reflects the stochastic nature of the neural weight distribution and assumes that, for each point in the ANN hyperspace, multiple estimators may produce the same value of $E_D$. Therefore, by applying Bayes' theorem, the posterior probability $P(w \mid D, \alpha, \beta, M)$ is introduced:

$$P(w \mid D, \alpha, \beta, M) = \frac{P(D \mid w, \beta, M) \cdot P(w \mid \alpha, M)}{P(D \mid \alpha, \beta, M)} \tag{14}$$

where $D$ is the input dataset, $M$ is the estimator mathematical model, $P(w \mid \alpha, M)$ is the prior probability, $P(D \mid w, \beta, M)$ is the likelihood function and $P(D \mid \alpha, \beta, M)$ is the normalization factor. More specifically, application of Eq. (14) requires the determination of the maximum posterior probability so that, for a given distribution of neural weights, the most probable model $M$ can be identified in each epoch. Globally, this corresponds to minimizing the objective function given by Eq. (13). The data distribution or likelihood function is given by the following expression [34]:

$$P(D \mid w, \beta, M) = \frac{e^{-\beta E_D(D|w,M)}}{Z_D} \quad ; \quad P(w \mid \alpha, M) = \frac{e^{-\alpha E_w(w|M)}}{Z_w} \tag{15}$$

where $Z_w = \left(\frac{2\pi}{\alpha}\right)^{N/2}$ and $Z_D = \left(\frac{2\pi}{\beta}\right)^{N/2}$. The normalization factor $P(D \mid \alpha, \beta, M)$ can be determined on the basis of Bayes' theorem:

$$P(\alpha, \beta \mid D, M) = \frac{P(D \mid \alpha, \beta, M) \cdot P(\alpha, \beta \mid M)}{P(D \mid M)} \tag{16}$$

Considering the general expression for the posterior probability:

$$P(w \mid D, \alpha, \beta, M) = \frac{e^{-F(w)}}{Z_F} \tag{17}$$

where $Z_F = \int d^3w\, e^{-F(w,\alpha,\beta)}$. Combining equations (14)–(17) yields the final posterior probability expression:

$$P(D \mid \alpha, \beta, M) = \frac{Z_F}{Z_D \cdot Z_w} \tag{18}$$

Determining the maximum posterior probability corresponds to minimizing the probability given in Eq. (18). The constants $\alpha$ and $\beta$ are obtained from $A = \log\big(P(D \mid \alpha, \beta, M)\big)$ by solving:

$$\frac{dA}{dk} = 0, \quad k \in \{\alpha, \beta\} \tag{19}$$

Further details of the derivation are given in [34, 16], resulting in the following:

$$\alpha = \frac{\gamma}{2E_w(w_p)} \quad ; \quad \beta = \frac{N - \gamma}{2E_D(w_p)} \quad ; \quad \gamma = K - \mathrm{tr}(H^{-1}) \tag{20}$$

where $0 \leq \gamma \leq K$, and $w_p$ denotes the most probable value of the neural weights obtained from the classical LM algorithm. The neural weights in the LM algorithm are updated in each epoch (iteration) according to the following formula:

$$W_{\text{new}} = W_{\text{old}} - \mu H^{-1} \cdot J \cdot e \tag{21}$$

In Eq. (20) and Eq. (21), the quantity $H \simeq 2\beta J^T \cdot J + 2\alpha I$ is the Hessian matrix, $J$ is the Jacobian of the first derivatives of the network error $e$ and $\mathrm{tr}$ is the trace operator.

A brief description of the LMBR algorithm can now be given. The procedure begins with loading the input data and the initial values of the constants $\alpha$, $\beta$, and $K$. After the data have been propagated through the network, the neural weight matrices $W$ are updated according to Eq. 21. Within the same epoch, the coefficients $\alpha$, $\beta$, and $\gamma$ are also updated, according to Eq. 20, and, at the end of the epoch, the convergence condition $|E_D| < \varepsilon \sim 10^{-4}, 10^{-5}$ is tested. If the condition is fulfilled, the entire cycle is terminated and the resulting model is saved. Otherwise, the cycle continues until convergence is achieved or the maximum number of epochs is reached.

### 3.4. Determining the optimal structure of artificial neural networks

Since the study involved two databases, special attention was paid to the fact that machine learning methods, including ANNs, are highly data-driven. Consequently, numerical experiments were performed separately to determine the optimal ANN structures for both

databases. The results showed that, for the Air Quality dataset, the optimal number of hidden neurons was 16 for the LMBR and LM networks and 11 for the LSTM network. In contrast, for the Beijing dataset, the optimal number of neurons in the hidden layer was 9 neurons for the LSTM network and 50 neurons for the LMBR and LM networks. In all cases, the input layer had a number of neurons equal to the number of predictors in the system, while the output layer had one neuron, since this was a regression task. To reduce the risk of overtraining, cross-validation was applied. The numerical experiments showed that 10-fold cross-validation provided the best results for the LSTM network on both databases, whereas 5-fold cross-validation provided the best results for the LMBR and LM networks. For all network configurations, the learning rate was set to the default value of 0.01, while the maximum number of epochs was 500. LM and LMBR operated in batch mode, while the LSTM network operated in semi-batch mode with a batch size of 80 samples.

## 4. Results and discussion

The results of the simulations are shown in Fig. 6 (Air Quality set) and Fig. 7 (Beijing PM2.5 set).

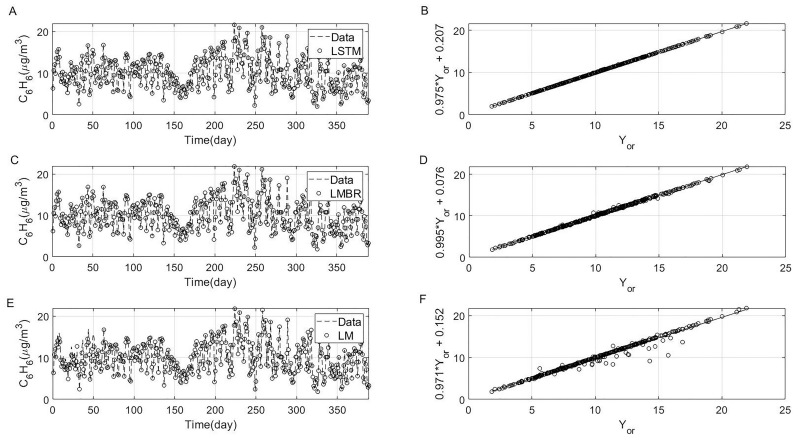


**Fig. 6.** Actual and simulated values of daily benzene concentrations for LSTM (A), LMBR (C) and LM (D) networks and regression plots: LSTM (B), LMBR (D) and LM (F)

To better visualize the differences in the simulation results, the corresponding values of the efficiency measures are listed in Table 5 and Table 6.

A visual inspection of the parameter values in Tables 5 and 6 shows that the LSTM networks perform efficiently in both cases, whereas LMBR and LM are competitive in the case of the Air Quality dataset. This can be explained by the presence of strong linear, i.e. monotonic, relationships between the benzene concentration and the predictors.

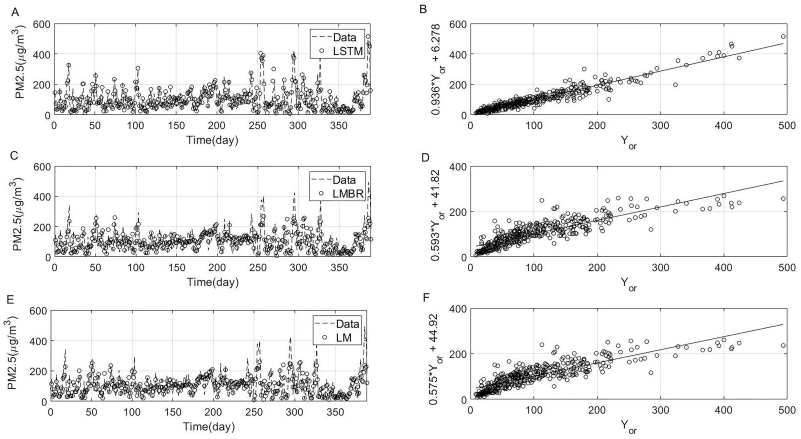


**Fig. 7.** Actual and simulated values of daily PM2.5 particle concentrations for LSTM (A), LMBR (C) and LM (D) networks and regression plots:LSTM (B), LMBR (D) and LM (F)

**Table 5.** Efficiency of ANNs in Air quality dataset

| **Model** | **RMSE** ($\mu$g/m$^3$) | **MAE** ($\mu$g/m$^3$) | **MAPE** (%) |
|---|---|---|---|
| LSTM | 0.11 | 0.09 | 1 |
| LMBR | 0.14 | 0.092 | 1 |
| LM | 0.57 | 0.2 | 2 |

**Table 6.** Efficiency of ANNs in Beijing PM2.5 dataset

| **Model** | **RMSE** ($\mu$g/m$^3$) | **MAE** ($\mu$g/m$^3$) | **MAPE** (%) |
|---|---|---|---|
| LSTM | 45.74 | 30.32 | 52 |
| LMBR | 64.94 | 42.83 | 72 |
| LM | 65.68 | 54.5 | 74.25 |

Based on the MAPE values reported in Tables 5 and 6, it can also be concluded that the LSTM networks are more robust in the case of the Beijing PM2.5 dataset, where monotonic relations are not dominant. The observed competitiveness of the LM/LMBR and LSTM networks across the two datasets provides an answer to research question RQ1. More importantly, the results highlight two key data characteristics that strongly influence model performance: sample size and monotonicity. This conclusion follows directly from the contrasting properties of the two datasets considered in this study.

Recent studies have examined the relationship between multilayer perceptron (MLP) networks, including LMBR/LM models and LSTM networks [66, 14, 50, 17]. The findings indicate that no universally optimal model exists, since performance depends largely on the structure of the dataset, the sequence length, and the processing of input attributes.

In general, LSTM networks tend to perform better in tasks involving sequential or temporal data, whereas MLP models can remain competitive when temporal dependencies are weak or when the dataset is relatively small ($<$10,000 samples). The results obtained in this study are consistent with these findings. Based on the RMSE, MAE, and MAPE values discussed previously, the performance of LMBR/LM networks is comparable to that of LSTM networks for the Air Quality dataset (9358 samples). However, for the Beijing dataset, where non-monotonicity relations dominate, LSTM networks achieve superior performance. Moreover, recent studies have shown that incorporating monotonic relationships between input variables can significantly improve the performance of ANNs, even when the available dataset is relatively small [63, 48].

The influence of monotonic relationships on the performance of ANNs is also indirectly confirmed by the Universal Approximation Theorem [23, 11]. According to this theorem, an ANN with a single hidden neural layer that uses a nonlinear transfer function can approximate any continuous function defined over a compact subset of $R^n$ to arbitrary accuracy. In the mathematical literature, it has been shown that every non-decreasing (monotonic) function is pointwise and uniformly continuous and that, in general, monotonic functions have a finite and countable set of discontinuities, often suggesting a strong tendency toward continuity [45, 2].

The significant influence of monotonicity relationships has also been confirmed by other authors. Building on the monotonicity requirement in traditional generalized linear models, the authors showed that introducing a monotonicity constraint in ANNs leads to improved predictive accuracy [46]. They achieved this by introducing a new set of hidden-layer transfer functions. This set consisted of three new zero-centered transfer functions combined with the original ReLU transfer function. This approach led to a fully connected neural layer with constrained monotonicity that can control the concavity and convexity properties of its output function. Accordingly, ANNs formed from these layers are called constrained monotone neural networks. Furthermore, the influence of the monotonicity property was also considered in a study related to the distribution of biological species as a function of environmental and climatic conditions modeled by a range of different variables [21]. As this field was dominated by traditional regression analysis models, linearity was an important assumption that was often not satisfied, and attempts to ease this assumption using the structured additive regression model (STAR) did not yield satisfactory results. Therefore, the authors introduced the concept of probability of finding a species at a given point in space and time. This probability is defined as the logistic transformation of the regression function: $f = \mathbf{B}\boldsymbol{\beta}$, where $\mathbf{B}$ is a spline in the form of a matrix $n \times m$ ($n$ – the number of samples and $m$ – the order of the spline), while $\boldsymbol{\beta}$ represents the corresponding coefficients in the form of a matrix $m \times 1$. The regression function was decomposed into a global component that takes into account the effects of the input variables and an additional component related to non-stationary effects and spatial and spatiotemporal autocorrelations. The model is estimated by minimizing the least-squares criterion, with the monotonicity constraint controlled by an additional sum of squared residuals. The results clearly showed that the fit with the monotonicity constraint was superior to the fit without this constraint. Another interesting approach that confirms the advantage of the monotonicity property is the Deep Lattice Network. Using a variant of these networks (Deep Lattice Cross Network), researchers predicted aerodynamic-force values with high accuracy [73]. This multi-purpose machine learning model for predict-

ing lift and drag coefficients was trained on the basis of fluid-dynamics simulation results. Excellent agreement was observed between the results of these models.

Tables 5 and 6 also show that LMBR networks are more efficient than LM networks for both datasets. The advantages of the LMBR algorithm have also been observed in other studies. Of particular interest is the sharp decline in the performance of all networks for the Beijing for the, where nonlinear and non-monotonic relationships between the input variables dominate. Based on these results, it can be concluded that the LSTM networks are more robust to the nonlinearity of the input data, as indicated by the parameter values in the tables above. These results can be considered in the context of the answer to research question RQ1.

In most studies, the theoretical analysis of the impact of nonlinear relationships in the system on LSTM networks has been limited to numerical experiments and comparisons with classical methods, because an analytical approach is often too complex or unavailable. For example, using a special type of modified LSTM network with recurrent feedback (OR-LSTM), researchers have shown that classical nonlinear system dynamics can be successfully reconstructed through nonlinear transfer functions and memory mechanisms [9]. Van der Pol and Duffing oscillators, as well as other nonlinear mechanical systems, were examined using OR-LSTM networks, and their typical characteristic behaviors, such as chaotic regimes, were successfully reconstructed. These findings are consistent with the results of the present study. Similarly, it has been shown that LSTM networks can successfully reconstruct high-dimensional chaotic systems such as the Lorenz 96 model and the Kuramoto–Sivashinsky system [60]. In this manner, researchers also examined nonlinear stochastic dynamical systems with noise, such as stochastic Van der Pol and Mackey–Glas oscillators [67]. It was shown that LSTM networks can track the evolution of the states of these dynamical models, not deterministically, but instead by modeling the probabilities of transitions between system states. This approach has proven to be more robust than classical methods. Similar results have been reported in several other studies [6, 32]

The advantage of LMBR networks on smaller datasets has also been observed in other studies. Using air pollution data from the Putrajaya area in Malaysia, the authors evaluated three different methods for air pollution analysis and compared their performance [37]. These methods were the autoregression moving average (ARIMA) model and 40 ANNs based on the LMBR and conjugate gradient algorithms. They forecasted the API (Air Pollutant Index) as the target variable as a function of various pollutants (CO, $O_3$, $PM_{10}$) and showed that the LMBR algorithm outperformed the other two methods based on the mean squared error (MSE) and MAPE values. The lowest MSE values were 19.43 for LMBR, 19.6 for the conjugate gradient method, and 96.74 for ARIMA, while the corresponding MAPE values were 8.57 for LMBR, 8.82 for the conjugate gradient method, and 23.44 for ARIMA.

The results shown in Fig. 1 and Fig. 3, obtained by the IVS technique, represent the answer to the second research question RQ2. This technique is usually used in other studies to filter the most influential variables in order to improve the performance of ML methods. Although the most significant variables are selected for the ML method in this way, some important relationships between the input variables are also eliminated in the process. In addition, it remains an open question whether the same level of efficiency could be achieved with further parameter tuning in the hyperspace using all input vari-

ables. In this study, IVS techniques were applied to detect seasonal trends in the input data. In both databases, temperature, relative humidity, and absolute humidity are identified as highly influential variables. The influence of air humidity was confirmed in another study in which the authors examined the possibility of removing volatile organic compounds from the air using non-thermal plasma technology (NTP) [33]. They examined the properties of benzene oxidation during dielectric barrier discharge in a plasma reactor, which consisted of two coaxial quartz tubes. The inner tube contained a silver wire with a diameter of 8 mm, which served as the inner electrode and was connected to a voltage source. The outer stainless steel electrode was grounded and attached to the wall of the outer electrode. A gas mixture consisting of benzene, nitrogen, and oxygen was introduced into the plasma formed between these two electrodes to simulate real industrial conditions. With controlled introduction of water vapor into this system, it was shown that benzene decomposition increased with relative humidity until it reached 60%, after which it began to decrease. The efficiency of benzene decomposition initially increases because benzene is oxidized by OH radicals formed through collisions between electrons and water molecules. In addition, OH radicals have a substantially higher oxidation potential than oxygen and other oxidants present in NTP plasma. At relative humidity values greater than 60%, the efficiency of benzene decomposition begins to decline, indicating that water vapor plays a dual role in the process.

A similar result was reported in another study in which researchers measured the concentration of air pollutants in the Mali Lošinj area in Croatia [19]. Using multisorbent sample tubes and the DDA technique, they determined the concentration of non-methane hydrocarbons. They showed that during the autumn period, when the concentration of OH radicals is low, the concentration of hydrocarbons is high, which indirectly implies the interaction of OH radicals and benzene. The most abundant hydrocarbons in their measurements were benzene, toluene, propane, and ethyne. An additional interesting finding was reported in the same study. Specifically, it was found that hydrocarbons with five or more carbon atoms, such as benzene, toluene, etc., are positively correlated with each other. This was also the case in the present study, where high statistically significant correlations of benzene with other hydrocarbons were also confirmed ($\rho = 0.77$). The most likely explanation for this phenomenon is traffic as a common source of NMHC and benzene, as shown in Fig. 5. This may also explain why NMHC was selected as an input variable, since the mrMR algorithm selects variables that are weakly correlated with each other and highly correlated with the output variable.

## 5. Limitations

In this work, input data filtering techniques were applied to increase the generality of the analysis, as they are independent of the AI methods to which they pass the selected variables. It is important to note that these techniques were used to improve data preprocessing and detect seasonal trends, although they are more commonly used to select the optimal set of input variables. In such cases, special attention must be paid to the fact that not all IVS methods are independent of the AI method to which they pass the selected variables. This is particularly important because, in addition to the ANN types examined in this study, there are other machine learning methods that can be combined with various IVS techniques.

## 6. Conclusion

A comparative analysis of LSTM networks and neural networks based on the LM algorithm was performed for air pollution prediction tasks. The results showed that ANNs based on the LM algorithm were competitive with LSTM networks when there were strong linear correlations between the target variable and the predictors. When the relationships between the variables were non-monotonic and nonlinear, the performance of all networks decreased significantly, although the LSTM network showed greater robustness than the other models. More broadly, this issue should also be examined in forecasting tasks, where LSTM networks are expected to have a substantial advantage due to their memory capabilities. This remains an important direction for future research. The results indicate that the structure of the input data remains one of the most influential factors affecting the performance of machine learning methods.

**Funding.** This research did not receive external funding.

**Conflicts of interest.** The authors declare no conflict of interest.

## References

1. Abirami, S., Chitra, P.: Regional air quality forecasting using spatiotemporal deep learning. Journal of Cleaner Production 283, 125341 (2021)
2. Apostol, T.M.: Mathematical Analysis. Addison-Wesley, Reading, MA, 2 edn. (1974)
3. Azan, A.N.A.M., Mototo, N.F.A.M.Z., Mah, P.J.W.: The comparison between arima and arfima model to forecast kijang emas (gold) prices in malaysia using mae, rmse and mape. Journal of Computing Research and Innovation 6(3), 22–33 (2021)
4. Batur, M., Babii, K.: The performance analysis of deep learning algorithms for modelling and forecasting the particulate matter (pm10) in the eastern part of turkey. In: IOP Conference Series: Earth and Environmental Science. vol. 1348, p. 012046. IOP Publishing (2024)
5. Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., Di Tommaso, S., Colangeli, C., Rosatelli, G., Di Carlo, P.: Recursive neural network model for analysis and forecast of pm10 and pm2. 5. Atmospheric Pollution Research 8(4), 652–659 (2017)
6. Bonassi, F., La Bella, A., Panzani, G., Farina, M., Scattolini, R.: Deep long-short term memory networks: Stability properties and experimental validation. In: 2023 European Control Conference (ECC). pp. 1–6. IEEE (2023)
7. Cateni, S., Colla, V., Vannucci, M.: Improving the stability of the variable selection with small datasets in classification and regression tasks. Neural Processing Letters 55(5), 5331–5356 (2023)
8. Chao, B., Guang Qiu, H.: Air pollution concentration fuzzy evaluation based on evidence theory and the k-nearest neighbor algorithm. Frontiers in Environmental Science 12, 1243962 (2024)
9. Chen, R., Jin, X., Laima, S., Huang, Y., Li, H.: Intelligent modeling of nonlinear dynamical systems by machine learning. International Journal of Non-Linear Mechanics 142, 103984 (2022)
10. Chen, S.: Beijing pm2.5 [dataset]. https://doi.org/10.24432/C5JS49 (2015), uCI Machine Learning Repository
11. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems 2(4), 303–314 (1989)

12. Dai, H., Huang, G., Wang, J., Zeng, H., Zhou, F.: Prediction of air pollutant concentration based on one-dimensional multi-scale cnn-lstm considering spatial-temporal characteristics: A case study of xi'an, china. Atmosphere 12(12), 1626 (2021)
13. Donnelly, A., Misstear, B., Broderick, B.: Real time air quality forecasting using integrated parametric and non-parametric regression techniques. Atmospheric Environment 103, 53–65 (2015)
14. Fei, X., Ye, M., Du, Z., Miao, H.: A comparative study of mlp and lstm neural networks for shale gas production prediction based on numerical simulation data. PLoS One 20(11), e0336782 (2025)
15. Gulati, S., Bansal, A., Pal, A., Mittal, N., Sharma, A., Gared, F.: Estimating pm2. 5 utilizing multiple linear regression and ann techniques. Scientific Reports 13(1), 22578 (2023)
16. Gull, S.F.: Developments in maximum entropy data analysis. In: Maximum Entropy and Bayesian Methods: Cambridge, England, 1988, pp. 53–71. Springer (1989)
17. Guo, Q., He, Z., Wang, Z.: Assessing the effectiveness of long short-term memory and artificial neural network in predicting daily ozone concentrations in liaocheng city. Scientific Reports 15, 6798 (2025)
18. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of machine learning research 3(Mar), 1157–1182 (2003)
19. Herjavić, G., Matasović, B., Arh, G., Kovač-Andrić, E.: Investigation of non-methane hydrocarbons at a central adriatic marine site mali lošinj, croatia. Atmosphere 11(6), 651 (2020)
20. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
21. Hofner, B., Müller, J., Hothorn, T.: Monotonicity-constrained species distribution models. Ecology 92(10), 1895–1901 (2011)
22. Hong, N., Liu, A., Zhu, P., Zhao, X., Guan, Y., Yang, M., Wang, H.: Modelling benzene series pollutants (btex) build-up loads on urban roads and their human health risks: Implications for stormwater reuse safety. Ecotoxicology and Environmental Safety 164, 234–242 (2018)
23. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural networks 2(5), 359–366 (1989)
24. Huang, C.J., Kuo, P.H.: A deep cnn-lstm model for particulate matter (pm2. 5) forecasting in smart cities. Sensors 18(7), 2220 (2018)
25. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial intelligence 97(1-2), 273–324 (1997)
26. Kovacs, B., Rusu-Both, R.: Neural networks based real-time indoor air quality monitoring and prediction iot system. In: 2025 29th International Conference on System Theory, Control and Computing (ICSTCC). pp. 812–817. IEEE (2025)
27. Lee, M.C., Chang, J.W., Hung, J.C., Chen, B.L.: Exploring the effectiveness of deep neural networks with technical analysis applied to stock market prediction. Computer Science and Information Systems 18(2), 401–418 (2021)
28. Li, T., Zhang, Q., Peng, Y., Guan, X., Li, L., Mu, J., Wang, X., Yin, X., Wang, Q.: Contributions of various driving factors to air pollution events: Interpretability analysis from machine learning perspective. Environment International 173, 107861 (2023)
29. Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., Chi, T.: Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environmental pollution 231, 997–1004 (2017)
30. Liu, Y., Wen, L., Lin, Z., Xu, C., Chen, Y., Li, Y.: Air quality historical correlation model based on time series. Scientific Reports 14(1), 22791 (2024)
31. Llamelo, C., Medina, R., Fajardo, A.: Performance of enhanced cnn-lstm prediction model for vehicle-mounted air quality monitoring. In: 2024 15th International Conference on Information and Communication Technology Convergence (ICTC). pp. 66–71. IEEE (2024)
32. Llerena Cana, J.P., Garcia Herrero, J., Molina Lopez, J.M.: Forecasting nonlinear systems with lstm: analysis and comparison with ekf. Sensors 21(5), 1805 (2021)

33. Ma, T., Zhao, Q., Liu, J., Zhong, F.: Study of humidity effect on benzene decomposition by the dielectric barrier discharge nonthermal plasma reactor. Plasma Science and Technology 18(6), 686 (2016)
34. MacKay, D.J.: Probable networks and plausible predictions-a review of practical bayesian methods for supervised neural networks. Network: computation in neural systems 6(3), 469 (1995)
35. Manisalidis, I., Stavropoulou, E., Stavropoulos, A., Bezirtzoglou, E.: Environmental and health impacts of air pollution: a review. Frontiers in public health 8, 14 (2020)
36. Méndez, M., Merayo, M.G., Núñez, M.: Long-term traffic flow forecasting using a hybrid cnn-bilstm model. Engineering Applications of Artificial Intelligence 121, 106041 (2023)
37. Mun, C.K., Abd Rahman, N.H., Ilias, I.C.: Performance of levenberg-marquardt neural network algorithm in air quality forecasting. Sains Malaysiana 51(8), 2645–2654 (2022)
38. Năstase, G., Șerban, A., Năstase, A.F., Dragomir, G., Brezeanu, A.I.: Air quality, primary air pollutants and ambient concentrations inventory for romania. Atmospheric Environment 184, 292–303 (2018)
39. Neo, E.X., Hasikin, K., Lai, K.W., Mokhtar, M.I., Azizan, M.M., Hizaddin, H.F., Razak, S.A., et al.: Artificial intelligence-assisted air quality monitoring for smart city management. PeerJ Computer Science 9, e1306 (2023)
40. Ngia, L.S., Sjoberg, J.: Efficient training of neural nets for nonlinear adaptive filtering using a recursive levenberg-marquardt algorithm. IEEE Transactions on Signal Processing 48(7), 1915–1927 (2002)
41. Prado-Rujas, I.I., Garcia-Dopico, A., Serrano, E., Córdoba, M.L., Pérez, M.S.: A multivariable sensor-agnostic framework for spatio-temporal air quality forecasting based on deep learning. Engineering Applications of Artificial Intelligence 127, 107271 (2024)
42. Radovic, M., Ghalwash, M., Filipovic, N., Obradovic, Z.: Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC bioinformatics 18(1), 9 (2017)
43. Ramírez-Gallego, S., Lastra, I., Martínez-Rego, D., Bolón-Canedo, V., Benítez, J.M., Herrera, F., Alonso-Betanzos, A.: Fast-mrmr: Fast minimum redundancy maximum relevance algorithm for high-dimensional big data. International Journal of Intelligent Systems 32(2), 134–152 (2017)
44. Ren, Q.: Air quality prediction based on lstm algorithm. In: Sixth International Conference on Electromechanical Control Technology and Transportation (ICECTT 2021). vol. 12081, pp. 1119–1127. SPIE (2022)
45. Rudin, W.: Principles of mathematical analysis. 3rd ed. (1976)
46. Runje, D., Shankaranarayana, S.M.: Constrained monotonic neural networks. In: International Conference on Machine Learning. pp. 29338–29353. PMLR (2023)
47. Sarigiannis, D.A., Karakitsios, S.P., Gotti, A., Papaloukas, C.L., Kassomenos, P.A., Pilidis, G.A.: Bayesian algorithm implementation in a real time exposure assessment model on benzene with calculation of associated cancer risks. Sensors 9(02), 731–755 (2009)
48. Schlieper, P., Dombrowski, M., Nguyen, A., Zanca, D., Eskofier, B.: Data-centric benchmarking of neural network architectures for the univariate time series forecasting task. Forecasting 6(3), 718–747 (2024)
49. Schürholz, D., Kubler, S., Zaslavsky, A.: Artificial intelligence-enabled context-aware air quality prediction for smart cities. Journal of Cleaner Production 271, 121941 (2020)
50. Seo, B., Yoon, Y., Lee, K.H., Cho, S.: Comparative analysis of ann and lstm prediction accuracy and cooling energy savings through ahu-dat control in an office building. Buildings 13(6), 1434 (2023)
51. Singh, S., Gupta, P.: Comparative study id3, cart and c4. 5 decision tree algorithm: a survey. International Journal of Advanced Information Science and Technology (IJAIST) 27(27), 97–103 (2014)

52. Stockwell, D.R., Peterson, A.T.: Effects of sample size on accuracy of species distribution models. Ecological modelling 148(1), 1–13 (2002)
53. Sun, W., Huang, C.: A hybrid air pollutant concentration prediction model combining secondary decomposition and sequence reconstruction. Environmental Pollution 266, 115216 (2020)
54. Tao, H., Jawad, A.H., Shather, A., Al-Khafaji, Z., Rashid, T.A., Ali, M., Al-Ansari, N., Marhoon, H.A., Shahid, S., Yaseen, Z.M.: Machine learning algorithms for high-resolution prediction of spatiotemporal distribution of air pollution from meteorological and soil parameters. Environment international 175, 107931 (2023)
55. Tijani, M.A., Nwiabu, N.D., Bennet, E.O.: An artificial neural network model for small and medium size data analysis using levenberg-marquardt optimization algorithm. Journal of Scientific and Engineering Studies (JSES) 10(6), 29–39 (2024)
56. Uwimana, E., Zhou, Y., Sall, N.M.: A short-term load demand forecasting: levenberg–marquardt (lm), bayesian regularization (br), and scaled conjugate gradient (scg) optimization algorithm analysis. The Journal of Supercomputing 81(1), 55 (2025)
57. Vabalas, A., Gowen, E., Poliakoff, E., Casson, A.J.: Machine learning algorithm validation with a limited sample size. PloS one 14(11), e0224365 (2019)
58. Ventura, P., Khodamoradi, M., Costa, R., Figueiras, P., Jardim-Gonçalves, R.: Real-time environmental monitoring in smart buildings using federated learning. Procedia Computer Science 263, 680–687 (2025)
59. Vito, S.: Air Quality [dataset] (2008), https://doi.org/10.24432/C59K5F, uCI Machine Learning Repository
60. Vlachas, P.R., Byeon, W., Wan, Z.Y., Sapsis, T.P., Koumoutsakos, P.: Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 474(2213) (2018)
61. Wang, W., Mao, W., Tong, X., Xu, G.: A novel recursive model based on a convolutional long short-term memory neural network for air pollution prediction. Remote Sensing 13(7), 1284 (2021)
62. Waseem, M., Lin, Z., Yang, L.: Data-driven load forecasting of air conditioners for demand response using levenberg–marquardt algorithm-based ann. Big Data and Cognitive Computing 3(3), 36 (2019)
63. Wu, X., Fu, D., Li, Z., Yang, T.: Graph-enhanced & monotonic embeddings: A novel approach to tabular data representation. Neurocomputing p. 132108 (2025)
64. Xayasouk, T., Lee, H., Lee, G.: Air pollution prediction using long short-term memory (lstm) and deep autoencoder (dae) models. Sustainability 12(6), 2570 (2020)
65. Xiao, F., Yang, M., Fan, H., Fan, G., Al-Qaness, M.A.: An improved deep learning model for predicting daily pm2. 5 concentration. Scientific reports 10(1), 20988 (2020)
66. Yahia, A.B., Kadir, I., Abdallaoui, A., Elazhari, K.: Architectural optimization of a multilayer perceptron (mlp) neural network enhanced by the levenberg–marquardt algorithm for predicting relative humidity: application to tangier, morocco. Water SA 51(3), 279–287 (2025)
67. Yeo, K., Melnyk, I.: Deep learning algorithm for data-driven simulation of noisy dynamical system. Journal of Computational Physics 376, 1212–1231 (2019)
68. Yonar, A., Yonar, H.: Modeling air pollution by integrating anfis and metaheuristic algorithms. Modeling Earth Systems and Environment 9(2), 1621–1631 (2023)
69. Yuan, H., Xu, G., Lv, T., Ao, X., Zhang, Y.: Pm2. 5 forecast based on a multiple attention long short-term memory (mat-lstm) neural networks. Analytical Letters 54(6), 935–946 (2021)
70. Zhang, B., Zhang, Y., Zhang, K., Zhang, Y., Ji, Y., Zhu, B., Liang, Z., Wang, H., Ge, X.: Machine learning assesses drivers of pm2. 5 air pollution trend in the tibetan plateau from 2015 to 2022. Science of the Total Environment 878, 163189 (2023)
71. Zhang, L., Liu, P., Zhao, L., Wang, G., Zhang, W., Liu, J.: Air quality predictions with a semi-supervised bidirectional lstm neural network. Atmospheric Pollution Research 12(1), 328–339 (2021)

72. Zhang, Y.L., Cao, F.: Fine particulate matter (pm2. 5) in china at a city level. Scientific reports 5(1), 14884 (2015)
73. Zhao, J., Zeng, L., Lin, A., Shao, X.: Deep learning prediction method for aerodynamic forces on morphing aircraft considering physical monotonicity. Advances in Aerodynamics 7(1), 7 (2025)
74. Zhou, Y., Chang, F.J., Chang, L.C., Kao, I.F., Wang, Y.S.: Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts. Journal of cleaner production 209, 134–145 (2019)

**Goran Keković** is an Associate Professor at Faculty of Information Technology, Alfa BK University. He received his PhD in 2008 from the School of Electrical Engineering, University of Belgrade. His early research focused on condensed matter theory, after which he shifted his interests toward biosignal analysis (EEG) and medical data processing, employing statistical methods and machine learning techniques. His current research interests include data processing and artificial intelligence.

**Rade Božović** received his M.Sc. degree in 2007, and Ph.D. degree in 2019 in Electrical Engineering from the School of Electrical Engineering, University of Belgrade. In 2020, he joined the ALFA BK University as an Assistant Professor. His research interests include wireless communication systems, cognitive radio, signal processing. Also, he works as technical manager at company CRONY, Belgrade, Serbia, specialized for wireless communication.

**Sonja Ketin** is a Professor at Academy of Technical and Art Applied Studies Belgrade (Serbia), Dpt. Railway College of Applied Sciences. She graduated from the Faculty of Technology and Metallurgy, University of Belgrade, in the field of chemical engineering. At the Faculty of Technical Sciences in Novi Sad, she enrolled in post-graduate studies in the field of Environmental Engineering (master's degree and doctoral dissertation).

**Vladimir Mikić** is an Assistant Professor at the Faculty of Information Technology, Alfa BK University. He received his Ph.D. in Information Technology from the same institution. His research interests include technology-enhanced learning, personalized e-learning systems, learning analytics, and artificial intelligence in education.

**Miloš Ilić** is an Assistant Professor at the Faculty of Information Technology, Alfa BK University. He earned his Ph.D. in Information Technology from the same faculty. His research focuses on intelligent tutoring systems, artificial intelligence, data-driven e-learning, and programming.

**Boban Vesin** is a Professor of IT and Information Systems at the University of South-Eastern Norway. His research focuses on engineering learning technologies using artificial intelligence, learning analytics, and personalized learning, with recent work extending toward explainable AI and its application in medical education and clinical decision support.

*Received: November 1, 2025; Accepted: May 20, 2026.*