



# Benzene Pollution Forecasting by Recurrent Neural Networks Tuned with Adapted Elk Herd Optimizer

Dejan Bulaja<sup>✉</sup>, Tamara Zivkovic<sup>✉</sup>, Milos Pavkovic<sup>✉</sup>, Vico Zeljkovic<sup>✉</sup>, Nikola Jovic<sup>✉</sup>, Branislav Radomirovic<sup>✉</sup>, Miodrag Zivkovic<sup>✉</sup>, Nebojsa Bacanin<sup>\*✉</sup>

Faculty of Informatics and Computing, Singidunum University, 11000 Belgrade, Serbia

\* Correspondence: Nebojsa Bacanin (nbacanin@singidunum.ac.rs)

**Received:** 01-25-2025

**Revised:** 02-25-2025

**Accepted:** 03-01-2025

**Citation:** D. Babic, T. Zivkovic, M. Pavkovic, V. Zeljkovic, N. Jovic, B. Radomirovic, M. Zivkovic and N. Bacanin, "Benzene pollution forecasting by recurrent neural networks tuned with adapted elk herd optimizer," *J. Ind Intell.*, vol. 3, no. 1, pp. 1–11, 2025. <https://doi.org/10.56578/jii030101>.



© 2025 by the author(s). Published by Acadlore Publishing Services Limited, Hong Kong. This article is available for free download and can be reused and cited, provided that the original published version is credited, under the CC BY 4.0 license.

**Abstract:** Benzene is a toxic airborne contaminant and a recognized cancer-causing agent that presents substantial health hazards even at minimal concentrations. The precise prediction of benzene concentrations is crucial for reducing exposure, guiding public health strategies, and ensuring adherence to environmental regulations. Because of benzene's high volatility and prevalence in metropolitan and industrial areas, its atmospheric levels can vary swiftly influenced by factors like vehicular exhaust, weather patterns, and manufacturing processes. Predictive models, especially those driven by machine learning algorithms and real-time data streams, serve as effective instruments for estimating benzene concentrations with notable precision. This research emphasizes the use of recurrent neural networks (RNNs) for this objective, acknowledging that careful selection and calibration of model hyperparameters are critical for optimal performance. Accordingly, this paper introduces a customized version of the elk herd optimization algorithm, employed to fine-tune RNNs and improve their overall efficiency. The proposed system was tested using real-world air quality datasets and demonstrated promising results for predicting benzene levels in the atmosphere.

**Keywords:** Air pollution; Benzene; Recurrent neural networks; Optimization; Hyperparameter optimization; Swarm intelligence; Elk herd optimization algorithm

## 1 Introduction

As a widespread issue and an environmental occurrence, atmospheric contamination represents a persistent and alarming danger to many industrialized and emerging countries around the world. Air contamination can be defined as the presence or mixture of chemical, physical, or biological substances that progressively alter the composition of the air. Predominantly triggered by human-related activities, air pollution arises from the combustion of fossil fuels, such as power plants, automobiles, and residential heating systems, as well as from natural phenomena like wildfires, volcanic outbursts, and similar events [1]. Numerous investigations have examined the consequences that atmospheric pollution imposes on human civilization: adverse health effects on individuals, including respiratory ailments, cardiovascular conditions, pulmonary disorders, and early mortality [2, 3]. Additionally, there are economic and social repercussions driven by escalating pollution levels, for instance, in Shanghai, declining property values have resulted in significant setbacks within the housing sector [4]. Some research further identifies a connection between population movements and the intensity of pollution. In the study [5], the authors studied patterns during China's Lunar New Year celebrations, where the Shanghai population can shrink by up to 60%, leading to a decrease in food preparation and transportation activities. Subsequently, this contributed to a measurable reduction in specific airborne contaminants. Environmental contamination also exerts a profound influence on natural ecosystems, with one of the most significant contributors being global warming. Altering climatic conditions, such as increasing global temperatures, can drastically harm and remodel habitats, potentially causing species extinction, melting polar ice, and elevated ocean levels [2]. Persistent air contamination further damages vegetation by altering leaf structures [6], while fauna suffers either direct contact or indirectly through the consumption of polluted nutritional sources.

Airborne pollutant particles are classified into two primary groups: PM10 and PM2.5, referring to particulate matter with diameters smaller than 10 and 2.5 micrometers, respectively. The pollutants most frequently monitored include sulfur dioxide (SO<sub>2</sub>), nitrogen oxides (NO<sub>x</sub>), carbon monoxide (CO), and ozone (O<sub>3</sub>) [7]. Although benzene

(C<sub>6</sub>H<sub>6</sub>) is not classified as particulate matter, it exists alongside these particles as an extremely hazardous atmospheric contaminant and a widely recognized cancer-causing substance. Benzene is a volatile organic compound (VOC) that can rapidly vaporize into the atmosphere and is commonly linked to pollution both indoor and open air settings. The main contributors to benzene emissions include automobile exhaust, particularly from gasoline-powered engines, discharges from petroleum refineries and chemical manufacturing plants, burning of fossil fuels like coal and oil, and natural events such as wildfires. Consequently, developing the ability to forecast the magnitude, specific pollutants, and temporal occurrence of benzene-related air contamination is of vital significance.

Artificial intelligence techniques, particularly machine learning (ML), are increasingly applied in atmospheric pollution prediction due to their capacity to capture intricate, nonlinear dependencies among environmental factors. Using extensive data sets, including meteorological conditions, vehicular flow, industrial operations and historical contamination records, frameworks such as random forests, support vector machines, and advanced neural networks can effectively estimate concentrations of pollutants like PM<sub>2.5</sub>, NO<sub>2</sub>, and benzene. These predictive models consistently surpass conventional statistical approaches, providing timely analyses and early alerts that assist in safeguarding public health and shaping environmental regulations. However, as described by Wolpert's no free lunch theorem [8], no singular ML methodology guarantees optimal results in all forecast scenarios.

An additional challenge arises from the heavy dependency of the ML model performance on the selection of hyperparameter settings. Therefore, precise calibration of these parameters is essential for achieving reliable predictions, a process widely acknowledged as an NP-hard optimization challenge, which cannot be resolved using standard deterministic algorithms and instead requires probabilistic or stochastic optimization strategies.

In this research, an innovative atmospheric pollution prediction system was introduced, built on recurrent neural networks (RNNs) [9]. These architectures are specifically designed to process time series air quality data and accurately estimate benzene contamination levels by identifying hidden sequential trends. To further enhance the resilience and precision of the model, a customized form of metaheuristics based on swarm intelligence was incorporated into the proposed system, namely the elk herd optimization algorithm (EHO) [10]. This refined EHO was used to autonomously adjust the hyperparameters of the RNN, ensuring that the network adapts swiftly to pollution prediction tasks while maintaining high detection accuracy.

This study was consequently motivated by three main goals:

- Creation of an upgraded version of the EHO algorithm, explicitly designed to surpass the limitations of the original approach and fine-tuned for tackling the benzene pollution forecasting problem.
- Design of an RNN-driven framework capable of capturing the complex temporal interdependencies associated with air contamination, enabling dependable predictions of the benzene level.
- Integration of the modified EHO mechanism into the RNN-based prediction model to execute hyperparameter tuning, with the aim of achieving superior performance tailored to the specific forecasting challenge.

The remainder of this paper is organized as follows. Section 2 provides an overview of pertinent studies on machine learning applications in air quality monitoring. In addition, it elaborates on hyperparameter tuning techniques and the RNN framework, emphasizing its function in processing sequential atmospheric datasets. Section 3 explains the fundamental principles of the EHO algorithm and introduces a modified variant of this metaheuristic approach. Subsequently, Section 4 details the experimental configuration, while Section 5 showcases the obtained outcomes. Finally, Section 6 reflects on the significance of the results and outlines potential avenues for future investigations.

## 2 Related Works

Beyond their harmful impacts on human well-being, benzene and its related aromatic compounds are highly reactive substances and serve as primary precursors in the formation of secondary organic aerosols (SOA) and ground-level ozone within the atmosphere. Photochemical transformations involving BTEX (benzene, toluene, ethylbenzene and xylene) are influenced by sunlight exposure and the availability of oxidative agents such as nitrogen oxides and various transient radicals like hydroxyl (OH), alkyl peroxides and hydrogen peroxide radicals [11, 12].

The complex and non-linear behavior of BTEX compounds and the seasonal fluctuations in their gas-particle phase transitions necessitate an interdisciplinary research perspective and sophisticated computational modeling approaches [13]. These advanced tools facilitate the exploration of environmental interconnections, broaden existing scientific understanding, and establish a foundation for future sustainable development. The authors [14] utilized multiple linear regression to estimate benzene concentrations based on independent variables such as other air pollutants and meteorological parameters, while the study [15] applied Bayesian hierarchical models to investigate benzene exposure linked to petrochemical industries, relating industrial discharges with pollution incidents and regional mortality disparities.

In general, researchers have used various methodologies, including extreme gradient boosting (XGBoost), Generalized AutoRegressive Conditional Heteroskedasticity (GARCH), artificial neural networks, and the light gradient boosting machine (LightGBM) to forecast concentrations of volatile organic compounds (VOCs), particulate matter (PM), polycyclic aromatic hydrocarbons (PAHs) or haze occurrences (e.g., [16–21]).

Machine learning techniques, designed to associate pollutant behavior with surrounding environmental factors, require customization for each specific scenario (dataset), a process inherently categorized as a nondeterministic polynomial-hard (NP-hard) problem. Performing this procedure manually is labor intensive and time-consuming. Furthermore, NP-hard tasks cannot be efficiently solved using conventional deterministic methods, as they would require excessive computational resources and impractical timeframes. In contrast, probabilistic algorithms—particularly those based on swarm intelligence metaheuristics—are capable of identifying near-optimal solutions within acceptable time limits.

Consequently, several studies have explored the utilization of metaheuristic strategies to refine and train artificial intelligence models or statistical regression techniques, in order to improve predictive performance and uncover deeper insights into the impacts of air quality on human well-being [22–25].

Prominent examples of successful hyperparameter calibration using metaheuristic algorithms span various sectors. These approaches have been implemented in healthcare [26], intelligent energy systems [27], software engineering [28], precision agriculture [29] and opinion mining [30]. The cybersecurity domain has also leveraged these techniques for tasks such as intrusion detection [31], insider threat identification [32], along with other applications like meteorological predictions [33] and ecology [34–36].

In this study, an enhanced version of the EHO algorithm was embedded in the RNN architecture, targeting the fine-tuning of RNN hyperparameters to improve the precision and adaptability of click fraud detection mechanisms.

## 2.1 Recurrent Neural Network (RNN)

RNNs [9] represent a subset of artificial neural models specifically engineered to process sequential datasets. Unlike conventional feedforward architectures, RNNs incorporate recursive connections that enable the retention of prior information, allowing the model to capture time-based dependencies. In numerous domains, such as natural language modeling and text generation, speech recognition and voice assistants, sentiment analysis to comprehend emotions hidden in textual documents, as well as temporal data prediction, RNNs have proven capable of learning intricate patterns that change over time. Practical implementations include financial modeling and forecasting of stock markets, weather prediction, tracking vital signs of patients in medical domain, video and music analysis and anomaly detection.

Within this framework, the RNN is structured to handle benzene records by preserving a hidden state that stores contextual knowledge from earlier time steps. At every moment  $t$ , the network processes an input  $X_t$  and refreshes its hidden state  $h_t$  by integrating the current input and the preceding hidden state  $h_{t-1}$ . This transition is governed by a nonlinear activation mechanism, specifically the hyperbolic tangent function ( $\tanh$ ). The update operation is mathematically expressed as follows:

$$h_t = \tanh(W_{xh}X_t + W_{hh}h_{t-1} + b_h) \quad (1)$$

In this context,  $W_{xh}$  denotes the weight matrix linked to the input data,  $W_{hh}$  represents the recurrent weight matrix associated with the hidden layer and  $b_h$  is the bias component. This recursive structure enables the network to retain information from earlier inputs, which is vital for identifying patterns in user click activity that may extend across several time intervals.

After updating the hidden state, the network computes the output  $Y_t$  by applying a linear transformation to the hidden representation:

$$Y_t = W_{hy}h_t \quad (2)$$

where,  $W_{hy}$  signifies the weight matrix that projects the hidden state into the output domain. The streamlined nature of this model ensures efficient training while preserving the ability to capture critical temporal characteristics embedded within the benzene pollution sequences. While modern studies often incorporate attention mechanisms to further boost predictive accuracy, this section concentrates on the core RNN architecture, which serves as the foundation for the prediction system.

## 3 Methods

This section initially presents the fundamental version of the EHO metaheuristic algorithm. It then highlights the limitations of the standard EHO approach and introduces an enhanced modification, which was subsequently utilized in the conducted experimental studies.

### 3.1 Original Elk Herd Optimization Algorithm

The elk herd optimization (EHO) algorithm [10]. is a relatively new swarm-based computational strategy inspired by the reproductive behavior observed in elk populations. The methodology is modeled on two core mating phases: rutting and calving.

During the rutting stage, the herd is fragmented into several family groups of varying sizes. This segregation occurs as dominant male elks compete to form groups containing multiple harems. In the calving stage, offspring are produced from the most dominant bulls and their respective harems. The original EHO model incorporates a control variable  $B_r$  denoting the initial bull proportion within the herd.

The algorithmic procedure starts by generating the initial elk group, structured as a population of individual bulls and harems. This population Elk herd is mathematically represented as a matrix defined in Eq. (3).

$$\mathbf{EH} = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_n^1 \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \cdots & \vdots \\ x_1^N & x_2^N & \cdots & x_n^N \end{bmatrix} \quad (3)$$

where, in  $n \times N$  the  $N$  shows the population size.

Eq. (4) illustrates the generation of each individual solution  $x^j$ .

$$x_i^j = lb_i + (ub_i - lb_i) \times U(0, 1) \quad (4)$$

where,  $ub$  represents the upper boundary and  $lb$  denotes the lower boundary of the search space.

The elk population is then arranged in ascending order based on their fitness scores. Family groups are formed according to the starting bull ratio  $B_r$ . The total number of families is determined by  $B = \lfloor B_r \times N \rfloor$ . The fitness assessment is used to select the top-performing males from the population. Individuals with the highest fitness values are designated bulls within  $B$ . These selected bulls then compete to form their harems.

Harem allocation is performed using a roulette-wheel selection strategy, where bulls in  $B$  are assigned harems proportionally based on their fitness relative to the total fitness sum. Each bull  $x^j$  is assigned a selection probability  $p_j$  calculated from its absolute fitness  $f(x^j)$  as shown in Eq. (5).

$$p_j = \frac{f(x^j)}{\sum_{k=1}^B f(x^k)} \quad (5)$$

During the calving phase, the offspring of each family is represented as  $x_i^j(t+1)$  produced by inheriting traits from both the father bull  $x^{h_j}$  and the mother harem  $x_i^j(t)$ . If the calf  $x_i(t+1)$  shares the same index  $i$  as the father, its generation follows Eq. (6).

$$x_i^j(t+1) = x_i^j(t) + \alpha \cdot (x_i^k(t) - x_i^j(t)) \quad (6)$$

where,  $\alpha$  is a random coefficient within  $[0, 1]$ , which influences the degree of inherited characteristics of a randomly selected elk  $x^k(t)$ . Higher  $\alpha$  values increase the randomness in the offspring, promoting diversity.

Alternatively, if the calf shares the index with the mother, the offspring  $x_i(t+1)$  is generated using both the mother  $x^j$  and the corresponding father  $x^{h_j}$  as formulated in Eq. (7).

$$x_i^j(t+1) = x_i^j(t) + \beta (x_i^{h_j}(t) - x_i^j(t)) + \gamma (x_i^r(t) - x_i^j(t)) \quad (7)$$

where,  $x_i^j(t+1)$  represents the  $i$ -th component of calf  $j$  in generation  $t+1$ ,  $h_j$  identifies the father bull of the  $j$ -th harem,  $r$  is the index of a randomly chosen bull from the population.

In line with natural elk behavior, there exists a possibility that the harem female mates with another bull if the dominant bull fails to guard her adequately. The parameters  $\gamma$  and  $\beta$  are random values within  $[0, 2]$  that determine the proportional influence of the father and random bull on the traits of the offspring.

At the end of each iteration, bulls, harems, and newborn calves are combined in all families. The population is then reclassified by fitness and the top performing individuals are preserved for the subsequent generation.

### 3.2 Modified EHO Approach

Although EHO is a relatively recent optimization technique that has demonstrated strong performance in various fields, there is room for improvement. Specifically, both the exploration and the exploitation phases of the original EHO framework present opportunities for refinement. The weakest point of the algorithm include susceptibility to premature convergence issue, as in some runs EHO can converge too fast to the local optimum, which is particularly expressed in high dimensional search domains. This may happen either due to insufficient exploration power in the early phases, or too aggressive exploitation of the attractive areas of the search space. Another weakness is sluggish converging speed in some runs, as EHO tries to avoid local optima pitfalls, especially if the population is lacking diversity. To address those obstacles, the current study introduces a dynamic modification of the algorithm designed to improve both components of the search process.

During the initial phase of execution (the first  $T/2$  iterations), the focus is on intensifying the exploration. The individual with the lowest fitness score (that is, the weakest candidate solution) is replaced by a newly generated individual created through a hybridization process. This new solution is formed by applying the uniform crossover technique - borrowed from the genetic algorithm (GA) methodology [37] - to a randomly selected pair of individuals from the population.

In the last half of the optimization process (the final  $T/2$ ), the algorithm shifts its emphasis to exploitation. In this stage, the poorest performing individual is replaced with a hybrid offspring generated by combining the elite (best performing) individual with a randomly chosen member of the population, again using the crossover operation. This improved version of the EHO algorithm is termed adaptive EHO (AEHO), and its detailed procedure is presented in Algorithm 1.

---

**Algorithm 1** ACOA metaheuristics pseudo-code Produce starting population  $P$  of  $N$  random

---

Produce starting population  $P$  of  $N$  random solutions

**while** ( $t < T$ ) **do**

**for** (every crayfish in  $P$ ) **do**

        Utilize original COA search process

**end for**

    Arrange individuals in  $P$  with respect to their fitness scores

**if** ( $t < T/2$ ) **then**

        Replace the poorest crayfish within  $P$  by a hybrid between a pair of arbitrary individuals, utilizing crossover mechanism [37].

**else**

        Replace the poorest crayfish within  $P$  by a hybrid between the best crayfish and an arbitrary crayfish, utilizing crossover mechanism [37].

**end if**

**end while**

**return** Crayfish with the best fitness score in  $P$

---

The complexity of metaheuristics techniques in terms of fitness function evaluations (FFE) is a principal metric for measuring their computational efficiency. FFEs represent the count of evaluations of the objective function in a single run. It allows a platform-independent way to perform side by side comparisons of metaheuristics approaches. Considering that this modification does not introduce additional FFEs, which are typically the most computationally intensive operations in metaheuristic algorithms, the proposed AEHO maintains the same computational complexity as the standard EHO in terms of FFEs.

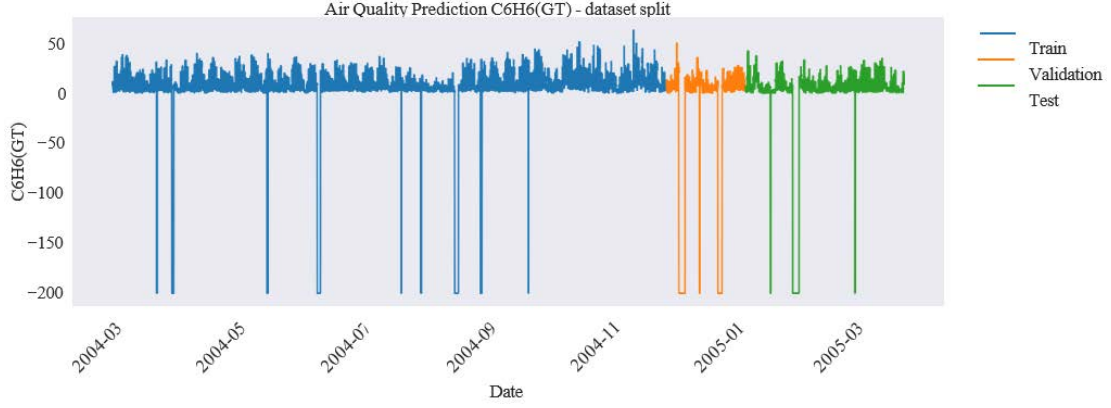
## 4 Experimental Setup

This research utilized openly accessible dataset from Kaggle for validation of the proposed benzene forecasting framework. The dataset is comprised of over 9300 entries of hourly mean readings gathered from a collection of five metal-oxide gas sensors integrated into an Air Quality Chemical Multisensor module [38]. This instrument was positioned outdoors at street level in a heavily contaminated urban zone located inside one Italian municipality. Recordings were captured between March 2004 and February 2005, spanning across an entire year, and represent the most extensive publicly accessible time series of in-situ air quality chemical sensor outputs. Confirmed ground truth data includes hourly averaged levels of carbon monoxide (CO), non-methane hydrocarbons (NMHC), benzene, total nitrogen oxides (NOx) and nitrogen dioxide ( $NO_2$ ). Benzene ( $C_6H_6$ ) was set as the target in this study. Dataset was split into 70%/10%/20% segments used for training, validation and testing, as outlined in Figure 1.

The capabilities of the proposed AEHO algorithm were compared to the collection of potent widely recognized metaheuristics, including baseline EHO [10], GA [37], particle swarm optimization (PSO) [39], bat algorithm (BA) [40] and COLSHADE [41]. Competing algorithms were individually crafted utilizing Python, employing the standard configuration settings for their control variables as recommended by their original developers. Every assessed optimizer was allotted a populace of 6 candidate solutions ( $N = 6$ ) and permitted 8 cycles ( $T = 8$ ) to carry out the optimization process. Due to the stochastic essence of metaheuristic methods, which inherently involve randomness, experiments were conducted over 30 independent executions. All considered algorithms were tasked with refining the models' efficacy via hyperparameters tuning. Table 1 delineates the assortment of optimized RNN configuration variables with their corresponding search intervals.

The performance of the generated RNN structures was evaluated using a standard set of regression KPIs [42], as outlined with Eqs. (8-11), including RMSE, MAE, MSE and  $R^2$ .

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (c_i - \hat{c}_i)^2} \quad (8)$$



**Figure 1.** Visualized data splits into train, validation and test subdivisions

$$\text{MAE} = \frac{1}{m} \sum_{i=1}^m |c_i - \hat{c}_i| \quad (9)$$

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^m (c_i - \hat{c}_i)^2 \quad (10)$$

$$R^2 = 1 - \frac{\sum_{i=1}^m (c_i - \hat{c}_i)^2}{\sum_{i=1}^m (c_i - \bar{c})^2} \quad (11)$$

A supplementary evaluation, known as the index of agreement (IoA) [43], was likewise monitored across the experiments, as it offers a more comprehensive insight into the RNN's performance. The IoA is calculated according to Eq. (12).

$$\text{IoA} = 1 - \frac{\sum_{i=1}^m (c_i - \hat{c}_i)^2}{\sum_{i=1}^m (|c_i - \bar{c}| + |c_i - \hat{c}_i|)^2} \quad (12)$$

Within the provided equations,  $c_i$  and  $\hat{c}_i$  correspond to the actual and predicted values of the  $i$ -th sample,  $\bar{c}$  denotes the mean score, whereas  $m$  denotes the length of entries. Throughout the conducted experiments, MSE was allocated as the fitness function that needs to be minimized, while  $R^2$  was employed to be the indicator function.

**Table 1.** RNN tuned parameters and their search intervals

Bound	Learning Rate	Dropout	Epoch Number	Cells within Layer	Count of Layers
Min	0.0001	0.05	100	100	1
Max	0.0100	0.20	300	250	2

## 5 Simulation Outcomes

Table 2 delineates the results of the fitness function optimization trials carried out over 30 separate runs, with the top score in each category highlighted in bold. The proposed AEHO demonstrated outstanding effectiveness, achieving the highest values for the best run, mean and median outcomes of 0.002265, 0.002413, and 0.002422, respectively. In this context, COLSHADE achieved the top score in the least favorable run, while it also demonstrated enhanced consistency in performance across independent runs, recording the lowest standard deviation and variance among the considered optimization algorithms.

Figure 2 illustrates a comparative side by side analysis of the consistency of the examined optimizers over multiple independent runs. The displayed violin plot indicates that the introduced AEHO is not the most reliable algorithm in terms of stability, as it is notably outperformed by several other metaheuristic strategies, like COLSHADE, the baseline EHO and BA. Nonetheless, although these alternatives demonstrated better uniformity in their outcomes, they failed to achieve the best overall performance, which was secured by AEHO. This outcome implies that the alternative methods are more prone to becoming trapped in local optima in comparison to the introduced AEHO. Within the same Figure 2, convergence diagrams of the fitness function (MSE) are also visualized, offering meaningful insight into



**Table 2.** RNN pollution forecasting outcomes in terms of fitness function (MSE)

Method	Best	Worst	Mean	Median	Std	Var
RNN-AEHO	<b>0.002265</b>	0.002563	<b>0.002413</b>	<b>0.002422</b>	0.000113	1.28E-08
RNN-EHO	0.002462	0.002720	0.002545	0.002507	0.000106	1.13E-08
RNN-GA	0.002355	0.002615	0.002463	0.002421	0.000095	9.04E-09
RNN-PSO	0.002457	0.002963	0.002625	0.002462	0.000208	4.33E-08
RNN-BA	0.002336	0.002517	0.002441	0.002459	0.000075	5.55E-09
RNN-COLSHADE	0.002328	<b>0.002435</b>	0.002374	0.002355	<b>0.000045</b>	<b>1.98E-09</b>

each method’s proficiency in avoiding local minima and converging toward more suitable areas of the search domain. It is evident that the introduced AEHO achieved the most favorable overall solution during the first round of its finest execution, surpassing other considered optimizers, which struggled to escape suboptimal regions. Furthermore, Figure 3 illustrates the box plots alongside the convergence curves of the  $R^2$ , offering additional perspective on the behavior and efficiency of the evaluated methods.

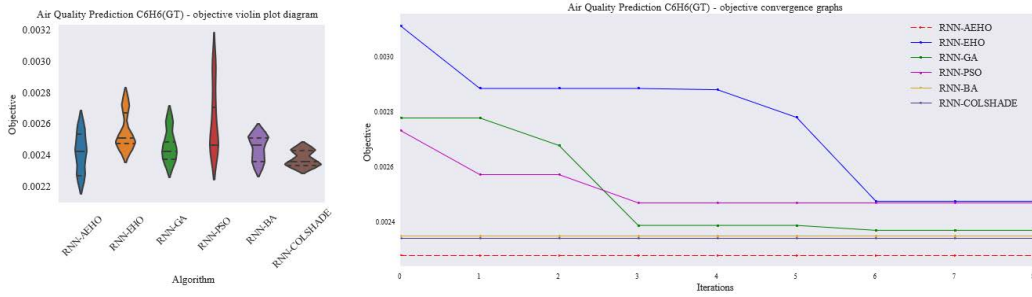
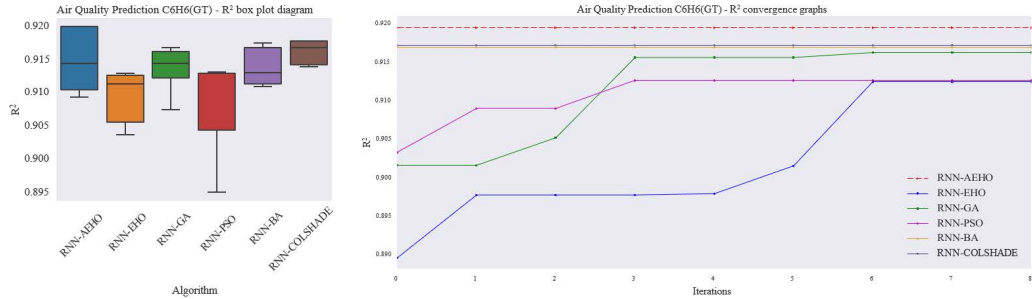
**Figure 2.** Objective function distribution and convergence diagrams**Figure 3.** Indicator function distribution and convergence diagrams

Table 3 offers a detailed overview of the comprehensive comparative assessment of performance indicators for the finest-performing RNN structures synthesized by each considered optimization technique. The proposed AEHO yielded an RNN configuration that achieved an impressive  $R^2$  of 0.918889, MAE of 0.018644, MSE of 0.002265, RMSE of 0.047594 and IoA of 0.978096. It is also evident that the remaining optimizers produced high-quality RNN models as well.

Figure 4 delineates the forecasts made by the finest-performing RNN tuned by suggested AEHO metaheuristics. Lastly, Table 4 depicts the hyperparameters’ configurations of the best RNN models synthesized with each observed optimization method, to support reproducibility of the simulations. The majority of the algorithms opted for structures with one hidden layer, except COLSHADE that determined RNN structure with two layers. Moreover, the RNNs with configurations listed in Table 4 attained the results delineated within Table 3.

## 6 Conclusion

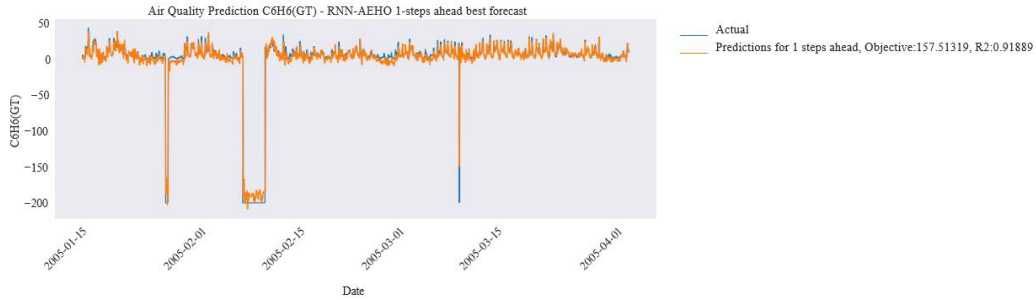
The spatiotemporal variability of air pollutant concentrations is remarkably dynamic, making air contamination a phenomenon of regional and global importance and a significant challenge for scientific investigation. The atmospheric behavior of harmful compounds depends on the intensity and nature of the emission sources, along

**Table 3.** RNN pollution forecasting outcomes in terms of detailed metrics

Method	$R^2$	MAE	MSE	RMSE	IoA
RNN-AEHO	0.918889	0.018644	0.002265	0.047594	0.978096
RNN-EHO	0.911850	0.019980	0.002462	0.049616	0.975955
RNN-GA	0.915653	0.018457	0.002355	0.048533	0.976746
RNN-PSO	0.912029	0.021413	0.002457	0.049565	0.975427
RNN-BA	0.916339	0.018608	0.002336	0.048336	0.977431
RNN-COLSHADE	0.916638	0.019480	0.002328	0.048249	0.977542

**Table 4.** RNN forecasting simulations best selected model hyperparameters

Method	LR	Dropout	E.No	L.No	L1	L2
RNN-AEHO	1.55e-03	5.17e-02	252	1	154	N/a
RNN-EHO	7.31e-04	1.98e-01	300	1	133	N/a
RNN-GA	5.67e-03	2.00e-01	300	1	100	N/a
RNN-PSO	1.00e-02	2.00e-01	300	1	100	N/a
RNN-BA	1.00e-02	2.00e-01	300	1	103	N/a
RNN-COLSHADE	1.00e-02	7.96e-02	276	2	157	228

**Figure 4.** RNN-AEHO best forecasted outcomes

with the meteorological conditions that influence their dispersion, chemical transformation, and eventual removal. Monitoring benzene, a highly volatile and carcinogenic compound, is essential for effective air quality assessment. Accurate prediction of benzene levels demands the deployment of advanced predictive frameworks.

This study explored the effectiveness of methodologies based on artificial intelligence, particularly those that employ RNN architectures, for estimating atmospheric benzene concentrations. To enhance the predictive performance of the RNN models, a modified version of the EHO algorithm was integrated for hyperparameter fine-tuning. The proposed system was evaluated using real-world data sets and produced encouraging results, with the top performing models achieving a MSE of only 0.002265 and a value  $R^2$  of 0.918889.

However, despite these promising findings, the research faced several limitations. The vast amount of available data imposed practical constraints on the size of the dataset utilized for training and evaluation phases. Furthermore, the computational demands of the optimization procedures restricted both the population sizes and the number of iterations allowed for the metaheuristic algorithms employed.

Recognizing significant potential for further enhancement, particularly since machine learning models remain insufficiently tested on complex environmental datasets, future work will involve benchmarking alternative machine learning models (both standard and metaheuristically optimized) against the framework developed in this study, while also applying them to other environmental datasets and challenges.

## Acknowledgements

This research was supported by the Science Fund of the Republic of Serbia, grant No. 7373, characterizing crises-caused air pollution alternations using an artificial intelligence-based framework (crAIRsis), and grant No. 7502, Intelligent Multi-Agent Control and Optimization applied to Green Buildings and Environmental Monitoring Drone Swarms (ECOSwarm).

## Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.



## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References

- [1] W. H. Organization, "Air pollution," [https://www.who.int/health-topics/air-pollution#tab=tab\\_1](https://www.who.int/health-topics/air-pollution#tab=tab_1), accessed: 21.01.2023.
- [2] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, "Environmental and health impacts of air pollution: A review," *Front. Public Health*, vol. 8, p. 14, 2020. <https://doi.org/10.3389/fpubh.2020.00014>
- [3] A. Haleem, A. H. Al-Obaidy, and S. Haleem, "Air quality assessment of some selected hospitals within Baghdad City," *Eng. Tech. J.*, vol. 37, no. 1C, pp. 59–63, 2019. <https://doi.org/10.30684/etj.37.1c.9>
- [4] G. Zou, Z. Lai, Y. Li, X. Liu, and W. Li, "Exploring the nonlinear impact of air pollution on housing prices: A machine learning approach," *Econo. Transp.*, vol. 31, p. 100272, 2022. <https://doi.org/10.1016/j.ecotra.2022.100272>
- [5] W. Zhou, W. Xu, Q. Wang, Y. Li, L. Lei, Y. Yang, Z. Zhang, P. Fu, Z. Wang, and Y. Sun, "Machine learning elucidates the impact of short-term emission changes on air pollution in Beijing," *Atmos. Environ.*, vol. 283, p. 119192, 2022. <https://doi.org/10.1016/j.atmosenv.2022.119192>
- [6] A. H. A. Al-Obaidy, I. M. Jasim, and A. R. A. Al-Kubaisi, "Air pollution effects in some plant leaves morphological and anatomical characteristics within Baghdad City. Iraq," *Eng. Tech. J.*, vol. 37, pp. 84–89, 2019. <https://doi.org/10.30684/etj.37.1c.13>
- [7] P. Perez and J. Reyes, "Prediction of particulate air pollution using neural techniques," *Neural Comput. Applic.*, vol. 10, no. 2, pp. 165–171, 2001. <https://doi.org/10.1007/s005210170008>
- [8] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997. <https://doi.org/10.1109/4235.585893>
- [9] L. Medsker and L. C. Jain, *Recurrent neural networks: Design and applications*. CRC press, 1999.
- [10] M. A. Al-Betar, A. Mohammed Awadallah, M. S. Braik, S. Makhadmeh, and I. A. Doush, "Elk herd optimizer: A novel nature-inspired metaheuristic algorithm," *Artif. Intell. Rev.*, vol. 57, no. 3, p. 48, 2024. <https://doi.org/10.1007/s10462-023-10680-4>
- [11] X. Cheng, Q. Chen, Y. Jie Li, Y. Zheng, K. Liao, and G. Huang, "Highly oxygenated organic molecules produced by the oxidation of benzene and toluene in a wide range of OH exposure and NO<sub>x</sub> conditions," *Atmospheric Chemistry and Physics*, vol. 21, no. 15, pp. 12 005–12 019, 2021. <https://doi.org/10.5194/acp-21-12005-2021>
- [12] Y. Deng, J. Li, Y. Li, R. Wu, and S. Xie, "Characteristics of volatile organic compounds, NO<sub>2</sub>, and effects on ozone formation at a site with high ozone level in Chengdu," *J. Environ. Sci.*, vol. 75, pp. 334–345, 2019. <https://doi.org/10.1016/j.jes.2018.05.004>
- [13] C. H. Whaley, E. Galarneau, P. A. Makar, M. D. Moran, and J. Zhang, "How much does traffic contribute to benzene and polycyclic aromatic hydrocarbon air pollution? Results from a high-resolution North American air quality model centred on Toronto, Canada," *Atmos. Chem. Phys.*, vol. 20, no. 5, pp. 2911–2925, 2020. <https://doi.org/10.5194/acp-20-2911-2020>
- [14] D. Galán-Madruga and P. Jesús García-Camero, "An optimized approach for estimating benzene in ambient air within an air quality monitoring network," *J. Environ. Sci.*, vol. 111, pp. 164–174, 2022. <https://doi.org/10.1016/j.jes.2021.03.005>
- [15] C. Jephcote and A. Mah, "Regional inequalities in benzene exposures across the European petrochemical industry: A Bayesian multilevel modelling approach," *Environ. Int.*, vol. 132, p. 104812, 2019. <https://doi.org/10.1016/j.envint.2019.05.006>
- [16] A. Stojić, D. Maletić, S. Stanišić Stojić, Z. Mijić, and A. Šoštarić, "Forecasting of VOC emissions from traffic and industry using classification and regression multivariate methods," *Sci. Total Environ.*, vol. 521, pp. 19–26, 2015. <https://doi.org/10.1016/j.scitotenv.2015.03.098>
- [17] M. Perišić, D. Maletić, S. S. Stojić, S. Rajšić, and A. Stojić, "Forecasting hourly particulate matter concentrations based on the advanced multivariate methods," *Int. J. Environ. Sci. Technol.*, vol. 14, no. 5, pp. 1047–1054, 2016. <https://doi.org/10.1007/s13762-016-1208-8>
- [18] A. Stojić, G. Jovanović, S. Stanišić, S. H. Romanić, A. Šoštarić, V. Udovičić, M. Perišić, and T. Milićević, "The PM<sub>2.5</sub>-bound polycyclic aromatic hydrocarbon behavior in indoor and outdoor environments, part II: Explainable prediction of benzo [a] pyrene levels," *Chemosphere*, vol. 289, p. 133154, 2022. <https://doi.org/10.1016/j.chemosphere.2021.133154>
- [19] M. Vega García and L. José Aznarte, "Shapley additive explanations for NO<sub>2</sub> forecasting," *Ecolo. Inform.*, vol. 56, p. 101039, 2020. <https://doi.org/10.1016/j.ecoinf.2019.101039>

- [20] H. Dai, G. Huang, H. Zeng, and F. Zhou, "PM2.5 volatility prediction by XGBoost-MLP based on GARCH models," *J. Cleaner Produ.*, vol. 356, p. 131898, 2022. <https://doi.org/10.1016/j.jclepro.2022.131898>
- [21] H. Dai, G. Huang, H. Zeng, and R. Yu, "Haze risk assessment based on improved PCA-MEE and ISPO-LightGBM model," *Systems*, vol. 10, no. 6, p. 263, 2022. <https://doi.org/10.3390/systems10060263>
- [22] T. Jônatas Belotti, S. Diego Castanho, N. Lilian Araujo, V. Lucas da Silva, T. A. Alves, S. Yara Tadano, L. Sergio Stevan, C. Fernanda Corrêa, and V. Hugo Siqueira, "Air pollution epidemiology: A simplified Generalized Linear Model approach optimized by bio-inspired metaheuristics," *Environ. Res.*, vol. 191, p. 110106, 2020. <https://doi.org/10.1016/j.envres.2020.110106>
- [23] A. Yonar and H. Yonar, "Modeling air pollution by integrating ANFIS and metaheuristic algorithms," *Model. Earth Syst. Environ.*, vol. 9, no. 2, pp. 1621–1631, 2022. <https://doi.org/10.1007/s40808-022-01573-6>
- [24] G. I. Drewil and R. J. Al-Bahadili, "Air pollution prediction using LSTM deep learning and metaheuristics algorithms," *Meas. Sensors*, vol. 24, p. 100546, 2022. <https://doi.org/10.1016/j.measen.2022.100546>
- [25] A. Sangeetha and T. Amudha, "A particle swarm optimization methodology to design an effective air quality monitoring network," *Environ. Dev. Sustain.*, vol. 23, no. 11, pp. 15 739–15 763, 2021. <https://doi.org/10.1007/s10668-021-01312-4>
- [26] L. Jovanovic, N. Bacanin, M. Zivkovic, M. Antonijevic, A. Petrovic, and T. Zivkovic, "Anomaly detection in ECG using recurrent networks optimized by modified metaheuristic algorithm," in *2023 31st Telecommunications Forum (TELFOR), Belgrade, Serbia*, 2023, pp. 1–4. <https://doi.org/10.1109/telfor59449.2023.10372802>
- [27] R. Damaševičius, L. Jovanovic, A. Petrovic, M. Zivkovic, N. Bacanin, D. Jovanovic, and M. Antonijevic, "Decomposition aided attention-based recurrent neural networks for multistep ahead time-series forecasting of renewable power generation," *PeerJ Com. Sci.*, vol. 10, p. e1795, 2024. <https://doi.org/10.7717/peerj-cs.1795>
- [28] S. J. Villonth, J. P. Villonth, T. Zivkovic, M. Zivkovic, L. Jovanovic, N. Bacanin, and J. Mani, "Adaboost optimized by Sinh Cosh algorithm for prediction of software defects," in *Communications in Computer and Information Science*. Springer Nature Switzerland, 2024, pp. 53–68. [https://doi.org/10.1007/978-3-031-82153-0\\_5](https://doi.org/10.1007/978-3-031-82153-0_5)
- [29] L. Jovanovic, M. Zivkovic, N. Bacanin, M. Dobrojevic, V. Simic, K. K. Sadasivuni, and E. B. Tirkolaei, "Evaluating the performance of metaheuristic-tuned weight agnostic neural networks for crop yield prediction," *Neural Compu. App.*, vol. 36, no. 24, pp. 14 727–14 756, 2024. <https://doi.org/10.1007/s00521-024-09850-4>
- [30] M. Dobrojevic, L. Jovanovic, L. Babic, M. Cajic, T. Zivkovic, M. Zivkovic, S. Muthusamy, M. Antonijevic, and N. Bacanin, "Cyberbullying sexism harassment identification by metaheuristics-tuned extreme gradient boosting," *Comp. Mat. Continua*, vol. 80, no. 3, pp. 4997–5027, 2024. <https://doi.org/10.32604/cmc.2024.054459>
- [31] P. Dakic, M. Zivkovic, L. Jovanovic, N. Bacanin, M. Antonijevic, J. Kaljevic, and V. Simic, "Intrusion detection using metaheuristic optimization within IoT/IIoT systems and software of autonomous vehicles," *Sci. Rep.*, vol. 14, no. 1, p. 22884, 2024. <https://doi.org/10.1038/s41598-024-73932-5>
- [32] D. Mladenovic, M. Antonijevic, L. Jovanovic, V. Simic, M. Zivkovic, N. Bacanin, T. Zivkovic, and J. Perisic, "Sentiment classification for insider threat identification using metaheuristic optimized machine learning classifiers," *Sci. Rep.*, vol. 14, no. 1, p. 25731, 2024. <https://doi.org/10.1038/s41598-024-77240-w>
- [33] L. Jovanovic, N. Bacanin, V. Simic, J. Mani, M. Zivkovic, and M. Sarac, "Optimizing machine learning for space weather forecasting and event classification using modified metaheuristics," *Soft Comput.*, vol. 28, no. 7-8, pp. 6383–6402, 2023. <https://doi.org/10.1007/s00500-023-09496-9>
- [34] H. Jalali, F. Keynia, F. Amirteimoury, and A. Heydari, "A short-term air pollutant concentration forecasting method based on a hybrid neural network and metaheuristic optimization algorithms," *Sustainability*, vol. 16, no. 11, p. 4829, 2024. <https://doi.org/10.3390/su16114829>
- [35] M. F. Javed, B. Siddiq, K. Onyelowe, W. A. Khan, and M. Khan, "Metaheuristic optimization algorithms-based prediction modeling for titanium dioxide-Assisted photocatalytic degradation of air contaminants," *Results Eng.*, vol. 23, p. 102637, 2024. <https://doi.org/10.1016/j.rineng.2024.102637>
- [36] Y. Qin, H. Fan, and A. Mohammed Al-Qaness, "PM2.5 prediction combined ANFIS with meta-heuristic optimization algorithms: A case study in Wuhan," in *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece*, 2024, pp. 2912–2916. <https://doi.org/10.1109/igarss53475.2024.10640686>
- [37] S. Mirjalili, *Evolutionary Algorithms and Neural Networks*. Springer International Publishing, 2019. <https://doi.org/10.1007/978-3-319-93025-1>
- [38] S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Di Francia, "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario," *Sen. Act. Che.*, vol. 129, no. 2, pp. 750–757, 2008. <https://doi.org/10.1016/j.snb.2007.09.060>
- [39] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95 - International Conference on Neural Networks, Perth, WA, Australia*, 1995, pp. 1942–1948. <https://doi.org/10.1109/icnn.1995.488968>
- [40] X. S. Yang and X. He, "Bat algorithm: Literature review and applications," *Int. J. Bio-Inspired Com.*, vol. 5,

no. 3, p. 141, 2013. <https://doi.org/10.1504/ijbic.2013.055093>

- [41] J. Gurrola-Ramos, A. Hernandez-Aguirre, and O. Dalmau-Cedeno, “COLSHADE for real-world single-objective constrained optimization problems,” in *2020 IEEE Congress on Evolutionary Computation (CEC)*, 2020, pp. 1–8. <https://doi.org/10.1109/cec48606.2020.9185583>
- [42] A. V. Tatachar, “Comparative assessment of regression models based on model evaluation metrics,” *Int. J. Inno. Tech. Exp. Eng.*, vol. 8, no. 9, pp. 853–860, 2021.
- [43] J. Cort Willmott, M. Scott Robeson, and K. Matsuura, “A refined index of model performance,” *Int. J. Cli.*, vol. 32, no. 13, pp. 2088–2094, 2011. <https://doi.org/10.1002/joc.2419>