MDPI

Article

Potential of Coupling Metaheuristics-Optimized-XGBoost and SHAP in Revealing PAHs Environmental Fate

Gordana Jovanovic [1,2,*], Mirjana Perisic [1,2], Nebojsa Bacanin [2], Miodrag Zivkovic [2], Svetlana Stanisic [2], Ivana Strumberger [2], Filip Alimpic [1] and Andreja Stojic [1,2]

1. Institute of Physics Belgrade, National Institute of the Republic of Serbia, University of Belgrade, 11000 Belgrade, Serbia; mirjana.perisic@ipb.ac.rs (M.P.); filip.alimpic@ipb.ac.rs (F.A.); andreja.stojic@ipb.ac.rs (A.S.)
2. Faculty of Informatics and Computing, Singidunum University, 11000 Belgrade, Serbia; nbacanin@singidunum.ac.rs (N.B.); mzivkovic@singidunum.ac.rs (M.Z.); istrumberger@singidunum.ac.rs (I.S.)
* Correspondence: gordana.vukovic@ipb.ac.rs

Abstract: Polycyclic aromatic hydrocarbons (PAHs) refer to a group of several hundred compounds, among which 16 are identified as priority pollutants, due to their adverse health effects, frequency of occurrence, and potential for human exposure. This study is focused on benzo(a)pyrene, being considered an indicator of exposure to a PAH carcinogenic mixture. For this purpose, we have applied the XGBoost model to a two-year database of pollutant concentrations and meteorological parameters, with the aim to identify the factors which were mostly associated with the observed benzo(a)pyrene concentrations and to describe types of environments that supported the interactions between benzo(a)pyrene and other polluting species. The pollutant data were collected at the energy industry center in Serbia, in the vicinity of coal mining areas and power stations, where the observed benzo(a)pyrene maximum concentration for a study period reached 43.7 ng m^{-3}. The metaheuristics algorithm has been used to optimize the XGBoost hyperparameters, and the results have been compared to the results of XGBoost models tuned by eight other cutting-edge metaheuristics algorithms. The best-produced model was later on interpreted by applying Shapley Additive exPlanations (SHAP). As indicated by mean absolute SHAP values, the temperature at the surface, arsenic, PM$_{10}$, and total nitrogen oxide (NOx) concentrations appear to be the major factors affecting benzo(a)pyrene concentrations and its environmental fate.

Keywords: machine learning; extreme gradient boosting; metaheuristics optimization; swarm intelligence; explainable artificial intelligence; sine cosine algorithm; benzo(a)pyrene

1. Introduction

Polycyclic aromatic hydrocarbons (PAHs) refer to a group of several hundred species with two to seven fused benzene rings, generated via incomplete combustion of organic substances, in the high temperature or pressure process. The majority of these polluting species are persistent, bioaccumulative, light sensitive, heat and corrosion resistant, and emitted from both natural and anthropogenic sources, with the latter being dominant in urban areas.

The concentrations of PAHs in the atmosphere are dependent on the number and quality of air pollutant emission sources, regional meteorological conditions, season, measurement site characteristics, as well as other factors which contribute to their dispersion and have an impact on atmospheric chemistry, dry or wet deposition, and finally, pollutant half-lives and their mutual interactions [1].

After being released, mostly as part of vehicle exhaust and emissions from biomass and fossil fuel burning, PAHs are distributed to all environmental compartments, adsorbed

to airborne particle matter, and deposited on terrestrial and water surfaces. Their concentrations are particularly high in the cold season, as a result of increased fossil fuel burning, reduced thermal and photo-decomposition, and a low planetary boundary layer. Apart from their common sources in urban areas, Hoffer et al. [2] reported that municipal incineration of plastic waste in urban areas emits up to 750 more PAHs than the combustion of dry firewood under the same conditions, and estimated that these emissions were dominated by 4–6 ring PAHs, which are up to 4100 times more toxic than the ones emitted from wood combustion.

In the atmosphere, PAHs are found in the gaseous phase or, more often, adsorbed onto suspended particles. The U.S. EPA has listed 16 compounds as the "priority PAHs" due to their adverse health effects, frequency of occurrence, and potential for human exposure. As regards their impacts on human health, PAHs have obtained significant attention due to the toxicity of low molecular weight species, being the most abundant in the gas phase, and the carcinogenic potential of heavy molecular weight compounds, being mostly particle-bound [3]. The smaller the particle size, the higher the share of carcinogenic PAHs, and thus fine aerosol fraction poses excessive risks to human health. In addition to this, PAHs contribute to the high mutagenicity and carcinogenicity of suspended particles through reactions with atmospheric oxidants, which result in the formation of secondary species [4].

In Serbia, the use of low-quality lignite coal is the major cause of low air quality. While domestic fuel burning (wood, coal, and gas) contributes to global $PM_{2.5}$ and PM_{10} emissions with 20% and 15%, respectively. Karagulian et al. [5] estimated that these contributions amount to 32% and 45% in Central and Eastern Europe, respectively. Since it has been recognized as an indicator of total exposure to carcinogenic PAHs, the benzo(a)pyrene (B[a]P) presence is regularly monitored. To prevent and reduce harmful effects on human health and the environment, a European Directive has set a target value of 1 ng m^{-3} for the total content of B[a]P in the PM_{10} fraction, averaged over a calendar year. In this study, we have used the pollutant data from Lazarevac, an energy industry center in the vicinity of the coal mining areas and power stations, where the observed B[a]P concentrations have occasionally reached 30 ng m^{-3}. In comparison to this, Elzein et al. [6] reported the $\sum$17-PAHs concentrations have ranged from 2.6 and 31.2 ng m^{-3}, and 8.4 to 42.9 ng m^{-3}, in Beijing and Delhi, respectively. Previous studies have confirmed that the residents of the coal mining regions face an incremental lifetime cancer risk which is significantly higher than the target value [7].

In this study, based on our previous research [8–15], we have applied a novel approach based on the XGBoost model to identify the factors which are mostly associated with the observed B[a]P concentrations and the environmental conditions which support and facilitate B[a]P level dynamics and its interactions with other polluting species. The XGBoost itself is an efficient model; nevertheless, its hyperparameters require tuning for each particular prediction task in order to achieve good performance on the observed dataset. Manual tuning of the hyperparameters is an extremely slow, time-consuming, and error-prone task that is considered to be NP-hard by nature. To address this, a variant of the SCA metaheuristics algorithm [16] has been used to optimize the XGBoost hyperparameters. Metaheuristics algorithms, being stochastic by nature, have been established as a common choice to tackle NP-hard challenges. By performing simulations for the sake of this research, the most promising metaheuristics algorithm was determined to be the sine cosine algorithm (SCA); in other words, it was selected empirically. Moreover, this paper also proposes a modified version of SCA, by hybridizing it with another algorithm, to cancel the limitations of the elementary SCA. Modified SCA was later utilized as a part of the machine learning framework, and tasked to tune the collection of the XGBoost hyperparameters for this problem. The results attained by the proposed model have been compared to the results of XGBoost models tuned by eight other cutting-edge metaheuristics algorithms. The best-produced model was later on interpreted by applying Shapley Additive exPlanations (SHAP).

2. Background

2.1. XGBoost

The XGBoost algorithm utilizes an adaptive training method for optimizing its objective function, where each step in the optimization process relies on the outcome of the previous step. The mathematical representation of the XGBoost model's objective function has been defined as follows:

$$F_o{}^i = \sum_{k=1}^{n} l\left(y_k, \hat{y}_k^{i-1} + f_i(x_k)\right) + R(f_i) + C, \tag{1}$$

where the t-th round loss is denoted by l, y_k and $\hat{y}_k$ denote target observed values and predictions, respectively; f_i are additive functions from the space of the regression trees, constant term is marked as C, while the model's regularization parameter R can be defined as:

$$R(f_i) = \gamma T_i + \frac{\lambda}{2} \sum_{j=1}^{T} w_j^2 \tag{2}$$

where T corresponds to the number of leaves in the tree, while w values denote the scores in the corresponding leaves [17].

In general, the complexity of the tree structure is inversely proportional to the values of the customization parameters γ and λ. The larger the values of these parameters, the simpler the tree structure becomes. The model's first and second derivatives, represented by g and h, respectively, are expressed as follows:

$$g_j = \partial_{\hat{y}_k^{i-1}} l\left(y_j, \hat{y}_k^{i-1}\right) \tag{3}$$

$$h_j = \partial_{\hat{y}_k^{i-1}}^2 l\left(y_j, \hat{y}_k^{i-1}\right) \tag{4}$$

The solution is obtained using the next two formulas:

$$w_j^* = -\frac{\sum g_t}{\sum h_t + \lambda} \tag{5}$$

$$F_o{}^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum g)^2}{\sum h + \lambda} + \gamma T, \tag{6}$$

where the loss function score is denoted by F_o^*, while the solution's weight values are marked by w_j^*.

2.2. Metaheuristics Optimization

NP-hard challenges, a common occurrence in computer science, necessitate the use of stochastic algorithms like metaheuristics because deterministic methods are impractical. Metaheuristics methods can be categorized into various families with respect to the natural phenomenon they utilize to steer the search process, such as evolution or insect behavior [18–20]. The most significant families of metaheuristic algorithms are nature-inspired methods (further divided into genetic algorithms and swarm intelligence), methods established upon certain physical phenomena (e.g., storm, gravity, electromagnetism), algorithms that imitate certain aspects of the human behavior (e.g., teaching and learning, or brainstorming, or actions taken on social media), and approaches based on mathematical laws to guide the search (e.g., trigonometric function oscillations).

Swarm intelligence is established upon the behavior manifested by massive groups comprised of relatively modest units; for example, insects or birds in swarms, that are able to manifest highly coordinated and sophisticated behavioral patterns while they hunt, feed, mate or migrate [21,22]. These algorithms have exhibited high efficiency in solving a variety of the real-world NP-hard challenges. Among many available algorithms, well-known samples are the particle swarm optimization (PSO) [23], the ant colony optimization

(ACO) [24], the firefly algorithm (FA) [25] and the bat algorithm (BA) [26,27]. Recently, a highly efficient group of algorithms were derived from the mathematical functions and their properties to steer the search procedure, including the sine-cosine algorithm (SCA) [16] and the arithmetic optimization algorithm (AOA) [28].

The reason why there is a range of population-based algorithms is due to the no-free-lunch theorem (NFL) [29]. The NFL discloses that no single method can be the best for all optimization problems. Consequently, one algorithm may excel in one task but entirely fail in another, leading to the need for diverse metaheuristics approaches and the requirement to choose a suitable method for each specific optimization challenge.

Recently, population-based algorithms have been a common choice to address numerous real-world problems. The application domains include predicting the number of COVID-19 cases [30,31], fog, cloud and cloud-edge computing systems organization [32–35], wireless sensors and IoT optimization [36–39], feature selection [40], image processing and classifying in medicine [41,42], global tuning challenges [43,44], credit card fraud identification [45,46], tracking and predicting air pollution [47,48], network and computer systems intrusion detection [49,50], and finally, tuning different ML structures [51–56].

2.3. Shapley Additive Explanations

To explain the obtained best-performing model, which is vital for understanding the process being modeled, we have applied the explainable artificial intelligence method SHAP. Avoiding the trade-off between accuracy and interpretability, SHAP provides a straightforward and meaningful interpretation of the model-derived decisions. It is based on Shapley values, calculated as a feature importance measure by a game-theory approach which provides an impact of features on individual predictions [57]. Apportioning the difference between the prediction and the average prediction among the features [58], Shapley values represent fairly distributed payouts among the cooperating players (features) depending on their contribution to the joint payout (prediction). SHAP assigns each feature importance as a measure of its contribution to a particular prediction and interprets the impact compared to a model's prediction if that feature took some baseline value (mean). This way, the method provides valuable insights into a model's behavior (1) overcoming the main drawback of inconsistency, (2) minimizing the possibility of underestimating the importance of a feature with a specific attribution value, and (3) capturing feature interaction effects based on a generalization of Shapley values and interpreting the model's global behavior while retaining local faithfulness [15,59]. The main challenges of the method include Shapley value computation and background data choice which can induce uncertain or unintuitive feature attributions.

In this study, we have used the relative SHAP values introduced by Stojic et al. [11] to gain an insight into relative relationships among feature attributions for each prediction. Relative SHAP values, defined as a share of absolute SHAP in total attributed importance of all features for the particular prediction, show the relative influence of a feature on the prediction.

We have used the Python SHAP implementation (SHAP Python package) and TreeExplainer [59] to obtain SHAP values that we have used to produce SHAP dependency plots, representing the change of feature importance over its value range.

3. Methods

3.1. Measurements Methods

For this study, the two-year daily concentrations (2018–2019; 645 observations) of particulate matter PM_{10}, its constituents (Pb, As, Cd, Ni, and B[a]P), and inorganic gaseous pollutants (NO, NO_2, NOx, and SO_2) were obtained from the regulatory air quality monitoring station (44°23′02″ N, 20°15′55″ E) in Lazarevac (Serbia). The meteorological data were obtained from the Global Data Assimilation System (GDAS1) with a $1.0° \times 1.0°$ spatial resolution for the longitude and latitude of the monitoring station.

The Sven Leckel SEQ 47/50-RV sampler was used for collecting 24-h samples of particulate matter. The mass concentrations of PM_{10}, Pb, As, Cd, Ni, and B[a]P were determined according to the standards EN 12341, EN 14902, and EN 15549, while the concentrations of NO, NO_2, NOx, and SO_2 were obtained in accordance with the sampling procedures standardized in EN 14211 and EN 14212.

PM_{10} were collected on quartz filters (Whatman QMA, 47 mm) daily, as described in the Standard SRPS EN 12341:2015 (Ambient air—Standard gravimetric measurement method for the determination of the PM_{10} or $PM_{2.5}$ mass concentration of suspended particulate matter, 2015). The filters were pre-fired to remove organic impurities, and the pre-conditioning of both non-exposed and loaded filters was performed prior to gravimetric measurements.

The concentrations of As, Cd, Cr, Ni, and Pb as PM_{10} constituents were determined as described in the EN 14902:2008/AC:2013 Standard (Ambient air quality—Standard method for the measurement of Pb, Cd, As, and Ni in the PM fraction of suspended particulate matter, 2008). Firstly, the CEN/TC 264 N779 procedure was applied for the extraction of the trace elements. In brief, the pieces of exposed quartz filters were treated with an acidic mixture of $HNO_3(c)/30\%$ H_2O_2/H_2O (3/2/5) using analytical grade reagents (Merck) and distilled/deionized water (MiliQ, 18.2 MΩ). The filters were digested in closed 100 mL Teflon vessels in the Anton Paar 3000 microwave accelerated reaction system and the concentrations of trace elements were determined by inductively coupled plasma–mass spectrometry (ICP-MS) (device Agilent 7500ce with Octopole Reaction System). Quality control and verification of the applied procedures for microwave digestion and multi-elemental trace analysis using ICP-MS was conducted using the 2783 NIST (National Institute of Standard and Technology, MD, USA) standard reference material analysis, containing a PM_{10} fraction of urban dust from a mixed industrial urban area of Vienna, collected on a polycarbonate membrane filter. The recovery values were within the satisfactory range of ±20% from the reference value.

B[a]P was determined by the procedure described in the SRPS ISO 12884:2010 Standard (Ambient air—Determination of total (gas and particle-phase) polycyclic aromatic hydrocarbons—Collection on sorbent-backed filters with gas chromatographic/mass spectrometric analyses, 2010). Parts of the exposed filters underwent a microwave extraction procedure with a solvent mixture of n-hexane and acetone (12.5 mL n-hexane: 12.5 mL acetone) according to EPA method 3546. After extraction, the solution volume was reduced by rotary evaporation under reduced pressure (55.6 kPa and 0.2 mL iso-octane) to 1 mL. Afterward, the n-hexane solution was reduced to 0.25 mL under a nitrogen stream. Known quantities of internal standards were added to estimate the method recovery. B[a]P was analyzed using gas chromatography coupled with a mass selective detector (Agilent GC 6890/5973 MSD) according to the EPA compendium method TO-13A with a DB-5 MS capillary column (30 m × 0.25 mm × 25 μm). The oven temperature program started at 70 °C (duration of 4 min) and ramped 8 °C min-1 to the end temperature of 310 °C (duration of 5 min). The solvent delay was 5 min and the run time was 46 min. The calibration curve was obtained by spiking seven different quantities of B[a]P, all with an R2 of the calibration curve above 0.995. Recovery values ranged from 85% to 110% for all the PAHs contained in the internal standard.

The samples were collected at the suburban site located in the energy industry center of Lazarevac, a municipality of Belgrade (Serbia), and a home to 60,000 residents. The sampling location, surrounded by residential areas and sports facilities, is exposed to mining pollutant sources and emissions from household coal and wood fireboxes. Additionally, the nearest coal mine Vreoci and regional power station are located around 5 km east and northeast, while the 80 square kilometers large coal mining area Kolubara, which supplies 75% of Serbia's electricity generation and largest state power plants Nikola Tesla A and B, are located around 10 km north and 30 km northwest of the sampling site, respectively.

3.2. Original Sine Cosine Algorithm

The algorithm, proposed by Mirjalili in 2016 [16], is based on the properties of elementary trigonometrical functions. SCA belongs to the group of population-based metaheuristics that starts each run by producing a set of arbitrary initial solutions within the scope of the search realm. The individual positions update following the swinging behavior of the sine and cosine functions over time. SCA conducts the exploration and exploitation mechanisms steered by the set of four arbitrary control parameters. The fundamental SCA search is mathematically defined by Equation (7):

$$X_i^{t+1} = \begin{cases} X_i^{t+1} = X_i^t + r_1 \cdot sin(r_2) \cdot |r_3 \cdot P_i^{*t} - X_i^t|, & r_4 < 0.5 \\ X_i^{t+1} = X_i^t + r_1 \cdot cos(r_2) \cdot |r_3 \cdot P_i^{*t} - X_i^t|, & r_4 \geq 0.5, \end{cases} \tag{7}$$

where X_i^t and X_i^{t+1} define the individual's position in *i*-th dimension in a pair of consecutive iterations t and $i + 1$, respectively, r_{1-4} are four generated above-mentioned control parameters, the P_i^* defines the target's position (the most recent estimation of the optimal solution) within *i*-th dimension. Additionally, the fresh values for r_{1-4} parameters are summoned for each component of each solution within the population.

3.3. Proposed Modified Sine Cosine Algorithm

The core implementation of SCA is considered to be an exceptional optimizer; however, like other metaheuristics methods, it also has some drawbacks. Testing using benchmark sets has shown that SCA is effective at exploring solutions, but lacks the ability to effectively exploit these solutions in the later stages of the process. This results in a limited exploration when the algorithm should be focusing on the most promising areas. In contrast, the firefly algorithm (FA) is known for its superior exploitation capability, as described by [25].

This manuscript suggests a hybrid solution by combining SCA and FA algorithms, aiming to profit from the advantages of both metaheuristics, aiming to cancel out each other's disadvantages. At the start of the execution, the solutions within the population will update according to the SCA search procedure, as described by Equation (7). However, in later stages, when it is necessary to narrow down and exploit the favorable regions of the search realm, the exploitation phase is backed up by employing the powerful FA search mechanism, defined by Equation (8):

$$X_i^{t+1} = X_i^t + \beta_0 \cdot e^{-\gamma r_{i,j}^2}(X_j^t - X_i^t) + \alpha^t(\kappa - 0.5) \tag{8}$$

where α represents the randomization variable, κ is an arbitrary value drawn from the Gaussian distribution. Finally, the space between solutions *i* and *j* is denoted as $r_{i,j}$.

A couple of new control parameters have been suggested to steer the alternation between the two search procedures in the later stage of the execution. The varying search *vs* control parameter is used to activate the combined search mode in the case where $t > vs$, when the suggested approach should alternate between SCA and FA search methods. Variable *vs* is initially set as *maxIter*/5, that has been determined empirically.

The second control parameter, named search mode *sm*, determines for every individual solution in the population whether to proceed with the SCA or FA search option. Each solution produces a random value *rnd* in range $[0, 1]$, and if $rnd < sm$ it will perform an SCA search, or otherwise continue with the FA search option. The value of this parameter is dynamically reduced over the iterations, giving an additional focus on a stronger FA search in the latter rounds. Initially, *sm* is set to 0.8, being reduced over time according to Equation (9).

$$sm_t = sm_{t-1} - (sm_{t-1}/10) \tag{9}$$

The hybrid algorithm is labeled hybrid self-adaptive SCA (HSA-SCA), and its pseudocode summarizing the most significant steps of the approach is provided by Algorithm 1.

Algorithm 1 Pseudocode of the HSA-SCA metaheuristics

Spawn a collection of starting solutions (X)
while $t < maxIter$ **do**
  validate each individual in terms of its fitness
  for each individual inside (X) **do**
    if $t < vs$ **then**
      Perform SCA search mechanism, provided by Equation (7)
    else
      if $rnd < sm$ **then**
        Perform SCA search mechanism, provided by Equation (7)
      else
        Perform FA search mechanism, provided by Equation (8)
      end if
    end if
  end for
end while
return The current fittest solution determined as the global optimum

---

4. Experimental Findings and Comparative Analysis

This section first provides insights into the dataset preprocessing, implementation technology, and evaluation metrics used to evaluate different tuned XGBoost models, followed by experimental setup, results, and comparative analysis. Finally, to validate improvements of devised hybrid metaheuristics over other baseline cutting-edge approaches, statistical tests were conducted, as suggested in the state-of-the-art AI literature [60].

4.1. Dataset Preprocessing, Implementation Technology and Evaluation Metrics

As already pointed out in Section 3, the employed dataset includes 645 observations. The challenge is formulated as a regression problem, where the feature with daily values for B[a]P was set as the target. Since the XGBoost is a tree-based method, scaling values, e.g., within the range $[0, 1]$, were not needed; therefore original measured values were used.

However, since the XGBoost requires training, the dataset was divided into train and test, where 70% of observations were used for training and 30% for testing. The same split was used for all metaheuristics considered for comparative analysis and the same pseudo-random number seed was employed, with the goal of establishing fair comparison conditions. It is noted that during the pre-experimentation, simulations with validation test were also conducted; however, improvements could not be achieved, and therefore it was decided to proceed with only training and testing data. Visual representation of the dataset split for the target variable is shown in Figure 1.
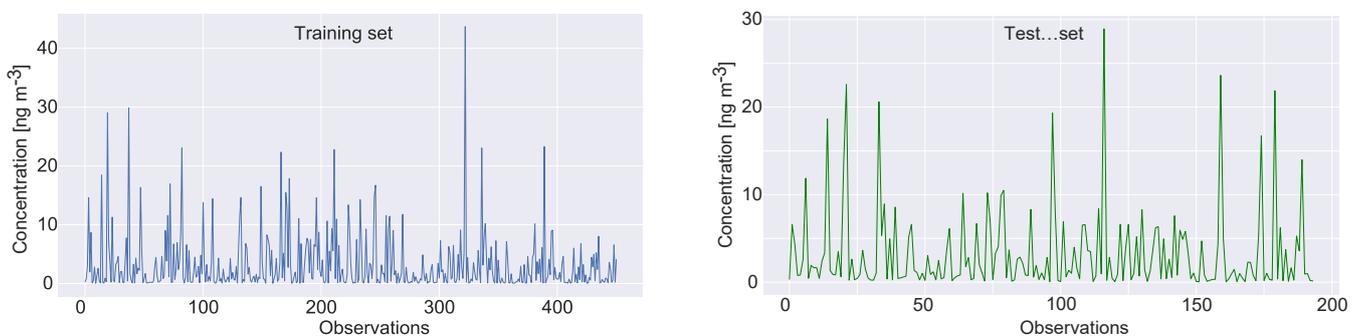

Figure 1. The Benzo(a)pyrene feature dataset split.

The analysis was conducted using daily concentrations and daily mean meteorological parameters, which have minimal to no autocorrelation in such a short period. Moreover, atmospheric processes relevant to air pollution dynamics usually occur within an hour, making autocorrelations even less prominent when using daily data.

The simulation environment, along with all methods, was implemented in Python using data science and ML libraries: *numpy, pandas, scikitlearn, xgboost, matplotlib, seaborn* and *shap*. Code snippets of the simulation framework along with the best generated XGBoost model by proposed HSA-SCA approach is available at the following URL: https://doi.org/10.5281/zenodo.7831739 (accessed on 25 February 2023).

The XGBoost model's experimental results have been evaluated by a set of traditional machine learning metrics, including mean squared error (MSE) defined by Equation (10), root mean squared error (RMSE) obtainable by Equation (11), mean absolute error (MAE) calculated by Equation (13), and the coefficient of determination (R2) described with Equation (13).

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (\hat{a}_i - a_i)^2 \tag{10}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{a}_i - a_i)^2} \tag{11}$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |\hat{a}_i - a_i| \tag{12}$$

$$R2 = 1 - \frac{\sum_{i=1}^{n} (a_i - \hat{a}_i)^2}{\sum_{i=1}^{n} (a_i - \bar{a})^2}, \tag{13}$$

where a_i and $\hat{a}_i$ represent arrays comprised of the observed values that are predicted, and predicted values, both with length N. This paper utilizes MSE as the fitness function that is required to be minimized.

Additionally, according to [61,62], the index of agreement (IA) can be an insightful statistical measure used to evaluate the performance of a model or forecast in predicting a particular event or phenomenon, as well as the metric for the best-generated models. The IA can be calculated as the ratio of the MSE and the potential error that is varying in range $[0, 1]$, where the value of 1 suggests perfect agreement, while the value of 0 suggests no match at all. The Eq. 14 shows how the IA value is obtained:

$$IA = 1 - \frac{\sum_{i=1}^{n} (a_i - \hat{a}_i)^2}{\sum_{i=1}^{n} (|\hat{a}_i - \bar{a}| + |a_i - \bar{a}|)^2}, 0 \leq IA \leq 1, \tag{14}$$

where a_i and $\hat{a}_i$ again denote arrays comprised of the observed and predicted values, and the $\bar{a}$ are average observed values.

4.2. Experimental Setup

The proposed HSA-SCA algorithm was tasked to optimize the XGBoost model for this particular dataset. The set of optimized XGBoost hyperparameters, accompanied by their corresponding search limits and variable types are provided as follows:

- learning rate (η), search limits: $[0.1, 0.9]$, continuous variable,
- *min_child_weight*, search limits: $[1, 10]$, continuous variable,
- subsample, search limits: $[0.01, 1]$, continuous variable,
- collsample_bytree, search limits: $[0.01, 1]$, continuous variable,
- max_depth, search limits: $[3, 10]$, integer variable and
- *gamma*, search limits: $[0, 0.8]$, continuous variable.

The parameter counts of the *softprob objective function* ('num_class':self.no_classes)were passed as a parameter to the XGBoost, while the remainder of the XGBoost parameters were set to XGBoost defaults during the simulations.

The suggested method has been implemented in the Python programming language, accompanied by the standard collection of Python libraries related to machine learnin including scipy, numpy, and pandas, while the XGBoost model was acquired from the scikit-learn package.

The proposed setup utilizes the solutions' encoding scheme that observes each solution as an array with length *l*, where *l* denotes the number of optimized hyperparameters. Hence, the value *l* has been set to six, to match the tuned parameters.

Aiming to validate the performance of the XGBoost model tuned by the suggested HSA-SCA algorithm, the achieved results were compared to the results attained by eight other contending powerful algorithms. The comparisons were executed with elementary SCA, genetic algorithm (GA) [63,64], PSO [23], ABC [65], FA [25], whale optimization algorithm (WOA) [66], harris hawks' optimization (HHO) [67] and chimp optimization algorithm (ChOA) [68]. Every contending algorithm has been implemented independently by the authors of this manuscript, with the control parameters set to the recommended values from their respective publications. Each algorithm has been given the same task, to tune the same set of XGBoost hyperparameters.

All metaheuristics algorithms were tested with 40 solutions in the population and 20 iterations per run, over the course of 20 separate runs. As previously noted, MSE was set as the fitness function that needs to be minimized.

4.3. Experimental Findings and Comparative Analysis

This section yields the attained experimental outcomes, for the observed HSA-SCA algorithm and other contenders. Tables 1 and 2 show the simulation outcomes with respect to the fitness function, accompanied by the detailed metrics achieved in the best individual run of each algorithm, where the best results in each category are marked in bold.

Table 1 shows detailed comparisons with respect to the fitness function (MSE) attained by XGBoost models optimized by the nine regarded algorithms (the proposed HSA-SCA and eight contenders). The results suggest that the HSA-SCA method displayed a supreme performance level, by achieving the best scores for key performance indicators (best, worst, mean, and median). FA scored the best results for standard deviation and variance, by delivering the most stable results. The second best result with respect to the best, worst, mean and median run values was also FA, followed by PSO and ChOA. The best attained score by the HSA-SCA XGBoost model was the MSE of 2.468293, and R^2 of 0.892845.

Table 1. Comparative results of the objective function (MSE) of the observed metaheuristics.

Method	HSA-SCA	SCA	GA	PSO	ABC	FA	WOA	HHO	ChOA
Best	**2.468293**	3.184137	3.192827	3.077008	3.153481	2.590850	3.206221	3.180909	3.129932
Worst	**2.893362**	3.605363	3.635218	3.773639	3.774347	2.918338	3.685749	3.584440	3.561904
Mean	**2.731538**	3.443475	3.413906	3.390799	3.369339	2.771516	3.466122	3.363900	3.379943
Median	**2.725915**	3.479927	3.423980	3.413375	3.306318	2.779478	3.472304	3.358337	3.371960
Std	0.114964	0.112574	0.144976	0.214997	0.157270	**0.097592**	0.109602	0.130492	0.112995
Var	0.013217	0.012673	0.021018	0.046224	0.024734	**0.009524**	0.012013	0.017028	0.012768

Table 2 presents the detailed metrics achieved in the best single run of all regarded algorithms. Once more, it can clearly be seen that the proposed HSA-SCA dominantly outperformed contenders in terms of all indicators—R^2, R, MSE, RMSE and IA, except MAE, where FA achieved the best score. Looking into the MSE that has been employed as the fitness function with a goal to minimize it, HSA-SCA exhibited superior performance with the score of 2.468293, in front of the FA that scored 2.590850, PSO in third place that achieved 3.077008, and ChOA finishing fourth with the score of 3.129932. In terms of the

IA metric, the proposed HSA-SCA was also superior, attaining the value of 0.970348. FA finished second, with IA value of 0.967438, while WOA was third with IA value of 0.961817.

Table 2. Detailed metrics for the best individual run of the observed metaheuristics.

	R^2	R	MAE	MSE	RMSE	IA
HSA-SCA	**0.892845**	**0.944905**	0.987179	**2.468293**	**1.571080**	**0.970348**
SCA	0.861769	0.928315	1.081976	3.184137	1.784415	0.960579
GA	0.861392	0.928112	1.056114	3.192827	1.786848	0.958925
PSO	0.866420	0.930817	1.096070	3.077008	1.754140	0.959703
ABC	0.863100	0.929032	1.085311	3.153481	1.775804	0.959175
FA	0.887525	0.942085	**0.981363**	2.590850	1.609612	0.967438
WOA	0.860810	0.927799	1.036537	3.206221	1.790592	0.961817
HHO	0.861909	0.928391	1.143855	3.180909	1.783510	0.961245
ChOA	0.864122	0.929582	1.048337	3.129932	1.769161	0.960678

Lastly, the sets of the XGBoost hyperparameters that have been established by the best run of every algorithm are provided within the Table 3. The best performing method, that was the proposed HSA-SCA, produced the XGBoost model with a learning rate of 0.535844, max_child_weight of 4.768378, a subsample of 0.920331, collsample_bytree of 0.899994, max_depth of 5, and gamma value of 0.037125. The XGBoost structure produced by the FA algorithm, that finished in second place, consisted of the learning rate of 0.473028, max_child_weight of 5.459757, a subsample value of 0.937122, collsample_bytree of 1.000000, max_depth value of 7, and finally, gamma value of 0.318114.

Table 3. Best solutions' determined XGBoost hyper-parameters set.

Method	l.r. (μ)	Max_child_weight	Subsample	Collsample_bytree	Max_depth	Gamma
HSA-SCA	0.535844	4.768378	0.920331	0.899994	5	0.037125
SCA	0.424673	6.830426	0.903051	1.000000	10	0.800000
GA	0.515505	1.239850	0.921408	1.000000	4	0.000000
PSO	0.469675	5.890036	0.966332	0.732996	7	0.514569
ABC	0.424717	6.756257	0.910125	0.797826	7	0.349785
FA	0.473028	5.459757	0.937122	1.000000	7	0.318114
WOA	0.518772	6.961853	0.976281	0.978017	4	0.408959
HHO	0.533272	6.254540	1.000000	1.000000	10	0.800000
ChOA	0.388340	2.995555	0.766726	1.000000	8	0.000000

The performed simulations are visualized in Figures 2 and 3, showing the convergence graphs, box plots, violin plots and swarm plots of all nine algorithms, for both fitness function (Figure 2) and R2 (Figure 3).

While looking into the Figures 2 and 3, it is possible to see that the HSA-SCA method exhibits a very fast converging speed, together with FA metaheuristics, that is a little bit faster at the beginning, but finishing behind HSA-SCA at the end. One can note that FA also exhibits the most stable results, closely followed by the WOA and SCA, as can be seen from the box plot diagrams. Finally, the swarm plots show the diversity of the population within the last round of execution of the best run of each algorithm. It is possible to conclude that all solutions of the HSA-SCA population were proximal to the optimum value.
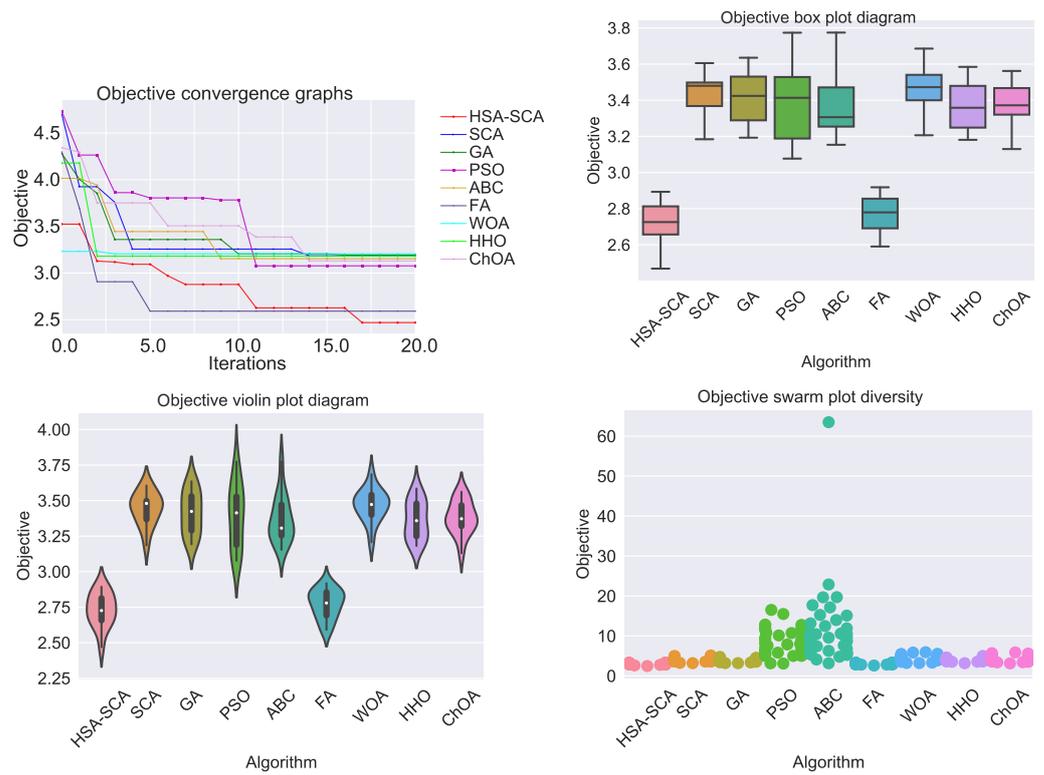
Figure 2. Visualized XGBoost results for all nine metaheuristics in terms of the convergence, box plot, violin diagrams, and swarm diversity plots for the fitness function (MSE).
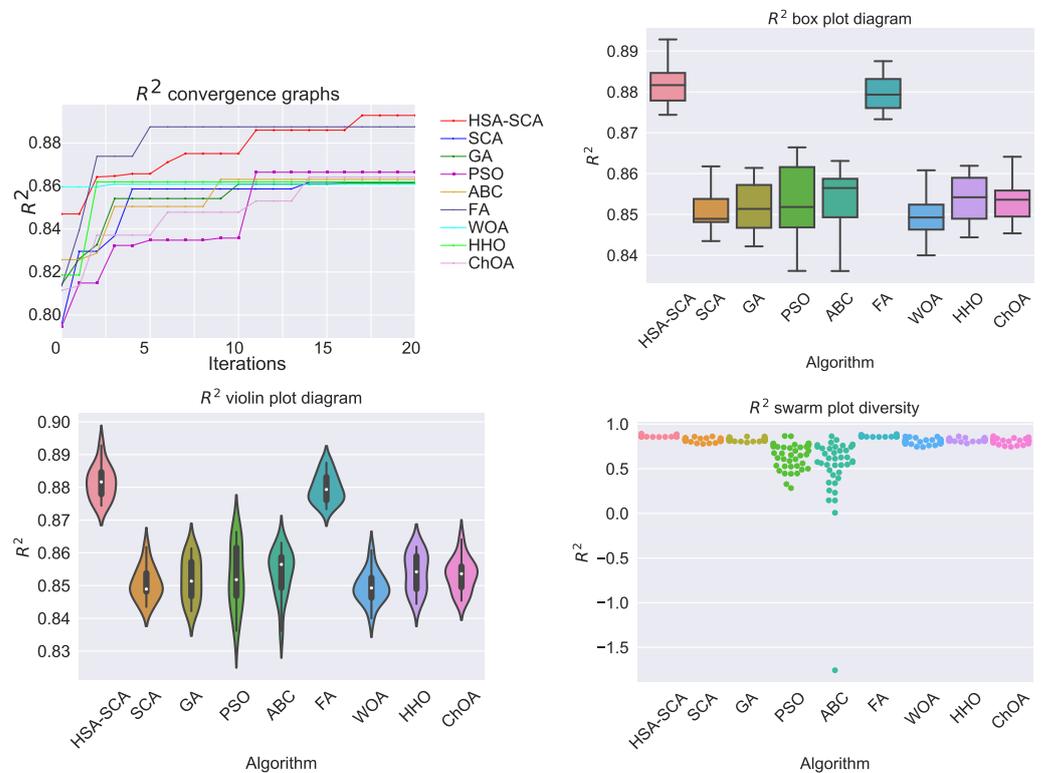

Figure 3. Visualized XGBoost results for all nine metaheuristics in terms of the convergence, box plot, violin diagrams, and swarm diversity plots for the R^2 indicator.

Figure 4 depicts the kernel density estimation (KDE), representing the estimation of the probability density function. It can be noted from these plots that the results originate from the normal distribution. Additionally, join plots of both fitness function (MSE) and R2 containing histograms for the two best algorithms (HSA-SCA and FA) are shown in Figure 5.
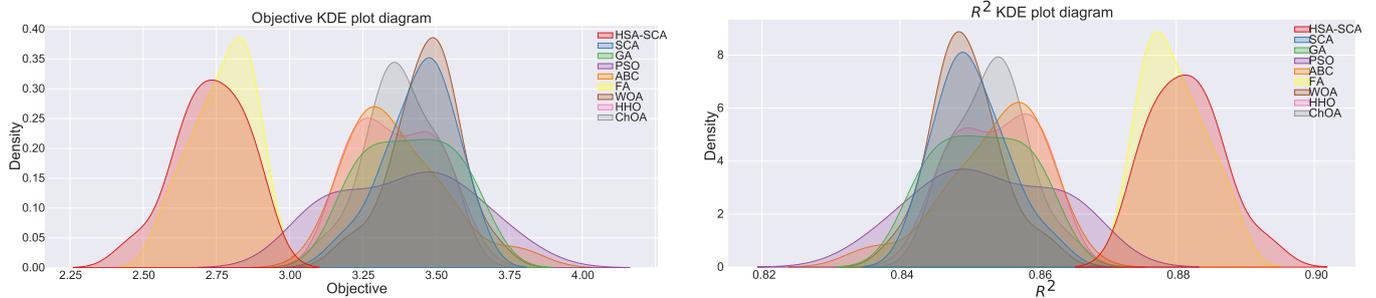
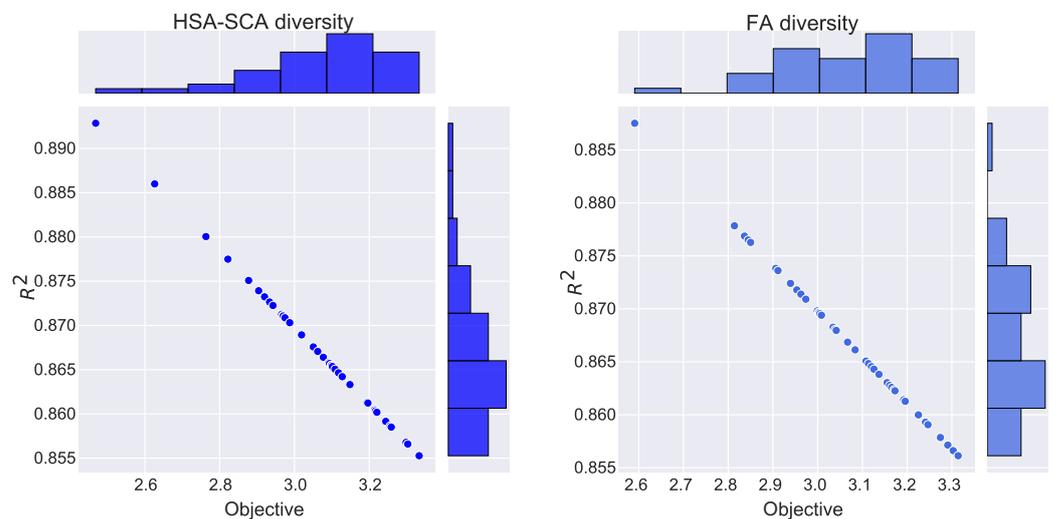**Figure 4.** KDE diagrams for MSE (**left**) and R^2 indicator (**right**).

Figure 5. Join plots with histograms of two best methods: HSA-SCA (**left**) and FA (**right**).

Finally, the visualizations of the best-predicted outcomes attained by the best-produced model by four best algorithms is shown in Figure 6. Once more, it can be concluded that the model optimized by the HSA-SCA algorithm produced the best predictions for the observed problem.
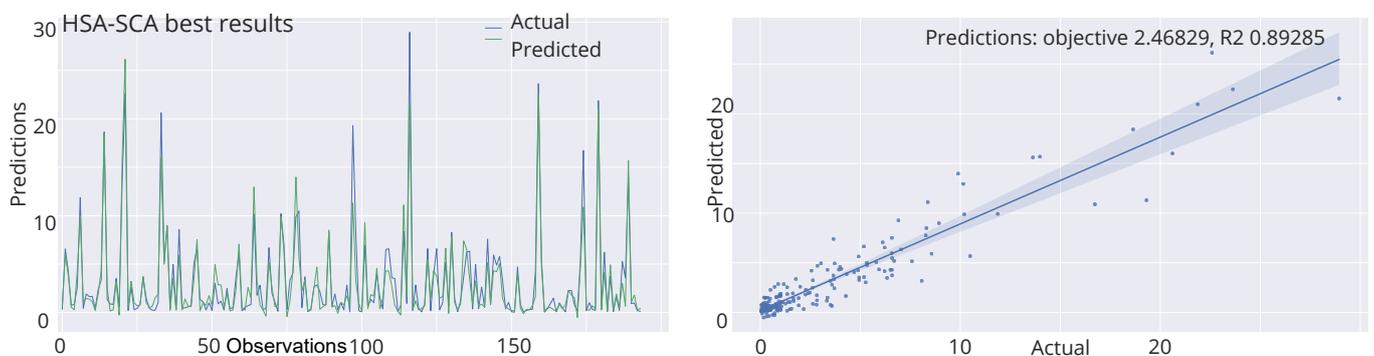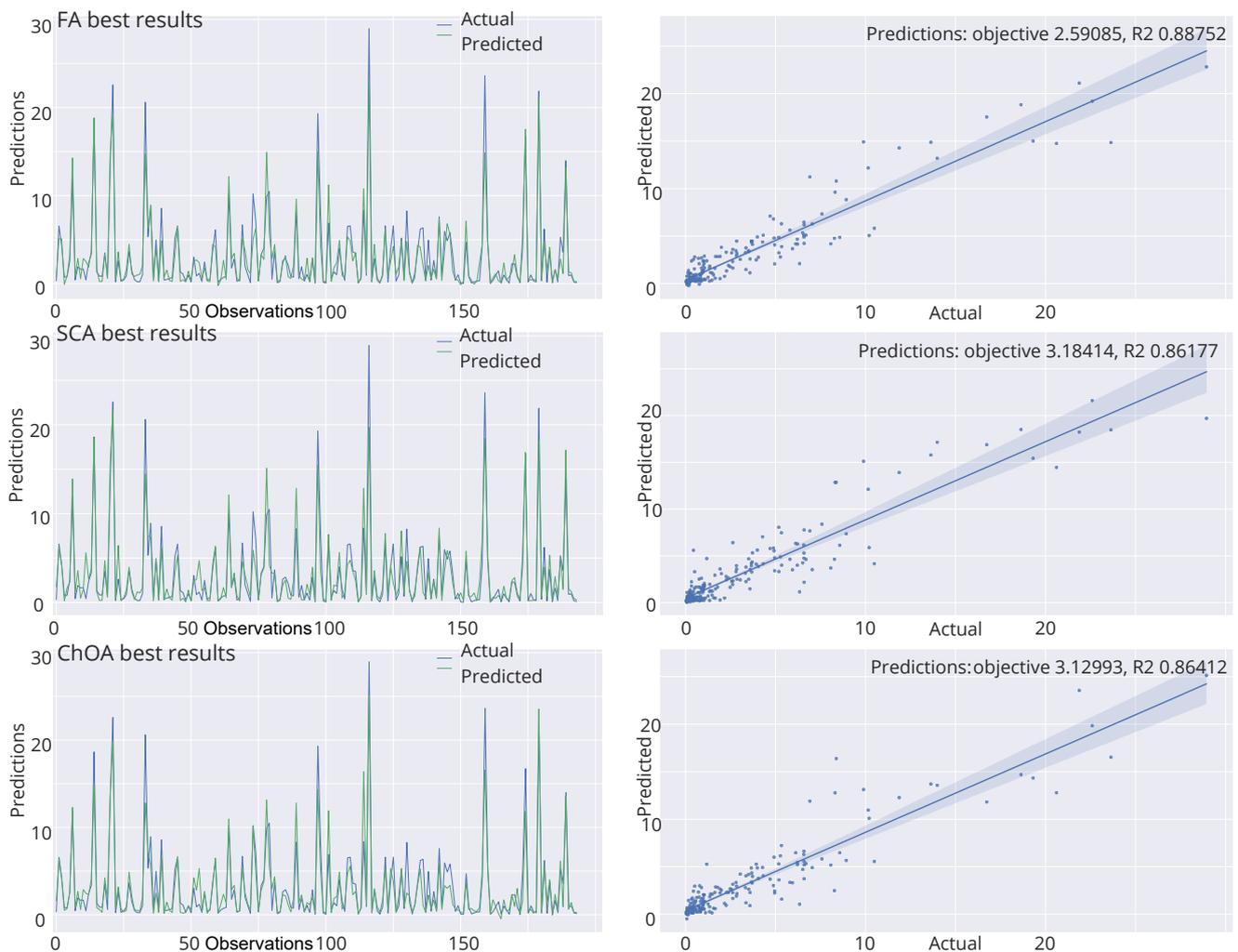

Figure 6. *Cont.*

Figure 6. Best-predicted outcomes by the best produced models of HSA-SCA, FA, SCA and ChOA algorithms.

4.4. Results Improvements Validation—Statistical Tests

To further evaluate the obtained simulation results and determine whether or not they are statistically significant, the best scores of each of 20 runs from each observed metaheruistics were gathered and inspected as a data series. At the beginning, it was necessary to decide what sort of statistical tests was suitable—parametric or non parametric. First, the safe usage of parametric tests is checked, by evaluation of the independence, normality, and homoscedasticity of the data variances [69]. The independence condition is satisfied, because every run of the metaheuristics algorithms begins by producing a collection of random individuals. Considering the second condition, homoscedasticity, Levene's test [70] was executed, and since the p-value of 0.65 was obtained in every case, it was safe to assume that the homoscedasticity requirement was also fulfilled.

The normality condition has been investigated by employing the Shapiro-Wilk single problem analysis [71]. Shapiro-Wilk p-values were independently calculated in terms of each of the observed methods. The established p-values for every algorithm were greater than 0.05, therefore it was safe to conclude that it was not possible to reject the H0 hypothesis for both $alpha = 0.05$ and $alpha = 0.1$. As a consequence, it was also possible to conclude that the observed values originated from the normal distribution. One could establish a similar conclusion by simply taking a look at the KDE plots shown in Figure 4. The Shapiro-Wilk test scores are provided within Table 4.

Table 4. Shapiro-Wilk scores for validating the normality condition.

Methods	HSA-SCA	SCA	GA	PSO	ABC	FA	WOA	HHO	ChOA
	0.362	0.351	0.198	0.145	0.312	0.304	0.342	0.263	0.347

After verifying that the normality requirement was fulfilled, it was safe to conclude that one can proceed by applying the parametric tests. This paper utilizes the paired-*t* test [72], which is frequently selected to evaluate metaheuristics methods [73]. Paired-*t* test can be utilized if it is possible to observe the set of data points as paired measurements, and the differences among the pairs follows a normal distribution. In other words, the variances between samples for every pair of algorithms are required to be normally distributed. To inspect this condition, the Shapiro-Wilk test was employed one more time, over the absolute differences between distributions of the proposed algorithm and other contending methods. The obtained Shapiro-Wilk *p*-values were in every instance greater than the threshold value $alpha = 0.05$, meaning that H0 hypothesis cannot be rejected, and the set of observed values originates from the normal distribution. Since this is the prerequisite for using the paired-*t* test, it is safe to use it and compare the proposed algorithm against each of the opposing methods. The summarized results of both Shapiro-Wilk *p*-values calculated as the prerequisite for the paired-*t* test, and the paired-*t* test itself, are provided in Table 5.

The results of the paired-*t* test show that the *p*-values was smaller then 0.05 for all algorithms. Accordingly, it can be concluded that the introduced HSA-SCA approach is significantly superior over all contenders for both thresholds $alpha = 0.1$ and $alpha = 0.05$.

Table 5. Shapiro-Wilk scores over the mean differences between two samples as prerequisite for paired-*t* test, accompanied by the paired-*t* test results

Methods vs. HSA-SCA	SCA	GA	PSO	ABC	FA	WOA	HHO	ChOA
Shapiro-Wilk	0.164	0.202	0.241	0.195	0.213	0.224	0.189	0.207
paired-*t* test	0.021	0.024	0.025	0.031	0.041	0.025	0.031	0.033

5. Discussion

The average annual B[a]P concentrations of 3.73 ng m^{-3} and 2.78 ng m^{-3} in 2018 and 2019, respectively (Table 6), significantly exceeded the European Directive set level of 1 ng m^{-3}. The maximum pollutant level reached 43.71 ng m^{-3} in the first year of the study period. At the same time, no values exceeded the critical threshold for the concentration of PM$_{10}$, As, Cd, Ni, and Pb, and inorganic gaseous pollutants.

Table 6. Descriptive statistics.

Year	Statistics	B[a]P [ng m^{-3}]	PM$_{10}$ [µg m^{-3}]	As [ng m^{-3}]	Cd [ng m^{-3}]	Ni [ng m^{-3}]	Pb [ng m^{-3}]	SO$_2$ [µg m^{-3}]	NO [µg m^{-3}]	NO$_2$ [µg m^{-3}]	NOx [µg m^{-3}]
2018	Average	3.73	36.18	1.56	0.41	2.78	4.35	10.82	10.63	17.88	29.96
	Minimum	0.07	10.2	0.5	0.05	1.5	2.5	2.09	0.5	0.5	0.5
	Maximum	43.71	179.7	11.2	27.1	19.6	30.7	25.02	101.49	98.5	189.71
	Median	1.15	29	0.5	0.2	1.5	2.5	12.45	4.61	14.33	19.3
2019	Average	2.78	33.01	1.38	0.2	2.26	4.26	17.21	4.98	11.91	19.43
	Minimum	0.03	4.9	0.5	0.05	1.5	2.5	5.7	1.1	1.75	6.1
	Maximum	23.63	180.7	13.6	1.9	28.5	40.6	66.3	91.2	48	146.81
	Median	1.07	23.3	0.5	0.1	1.5	2.5	15.45	2.75	10.45	14.6

As indicated by the mean absolute SHAP values, the temperature at surface (TMPS), As, PM$_{10}$, and total nitrogen oxide (NOx) concentrations appear to be major factors for governing B[a]P environmental fate (Table 7). In addition, the most important variables also include NO, SO$_2$, Pb, and Cd concentrations, as well as the temperature at 2 m (T02M) and

momentum flux intensity (MOFI)m have been shown to affect B[a]P dynamics. However, for this paper, we will focus on the aforementioned four.

Table 7. SHAP values.

Parameter	TMPS	As	PM$_{10}$	NOx	NO	SO$_2$	TO2M	Pb	MOFI	LIB4	SHIF	LHTF
Absolute	1.17	0.906	0.796	0.608	0.321	0.247	0.192	0.161	0.158	0.158	0.141	0.131
Relative [%]	22.7	14.36	13.75	9.99	4.99	4.25	3.98	3.38	2.45	2.99	2.18	2.33

5.1. Temperature at Surface

In this study, the temperature at the surface (TMPS) was estimated to be the most important parameter responsible for the B[a]P concentration increase of 1.17 ng m^{-3} on average, while mutual interrelations between TMPS and other studied parameters define three types of environmental conditions being responsible for shaping B[a]P levels. As a high-molecular-weight PAH, B[a]P is dominantly particle-bound in the atmosphere. The B[a]P partition between gas and particles is enhanced during colder months due to low temperature and high atmospheric pressure, which cause intense descending air movements and dry deposition of organic compounds [74]. Additionally, previous studies have shown that higher organic carbon content of particles in the cold season negatively affects the immobilization and biodegradation of PAHs [75], while high temperatures and light intensity in warm months enable both their photo- and biodegradation.

The first type of environment resulting in the increase of B[a]P concentrations up to 3.4 ng m^{-3} (Figure 7), was characterized by medium to low PM$_{10}$, B[a]P, As, Cd, and Ni levels (35.2 μg m^{-3}, and 3.4, 1.4, 0.3, and 2.5 ng m^{-3} on average, respectively), medium to high NO and NOx concentrations (6.9 and 23.4 μg m^{-3} on average, respectively), and meteorological parameters registered in a wide range of values. The observed constancy of the conditions suggests that this environment type might be related to anthropogenic sources, such as traffic and off-road vehicles.
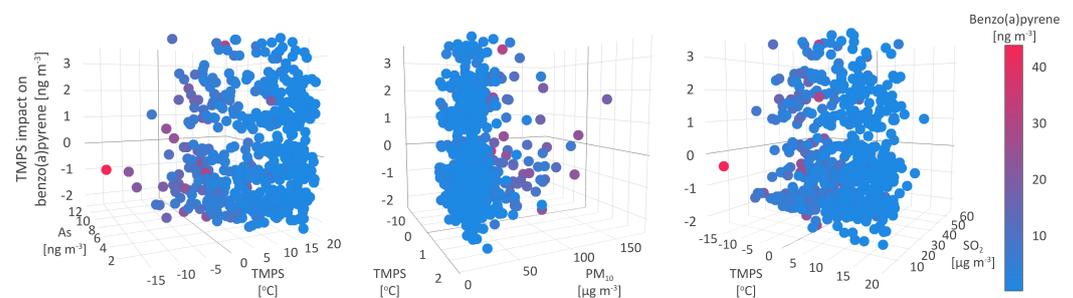

Figure 7. Temperature at surface impact on benzo(a)pyrene.

In the second type of environment, TMPS was ambivalently related to the B[a]P concentrations, leading both to their decrease by up to −1 ng m^{-3} and the increase by up to 0.7 ng m^{-3}. Compared to the first type, the second type of environment was characterized by lower B[a]P, As, Cd, Ni, Pb, PM$_{10}$, NO$_2$, and SO$_2$ (2.0, 1.1, 0.2, 1.8, 3.8 ng m^{-3}, and 32.8, 13.8, and 13.4 μg m^{-3}, respectively) and higher NO and NOx (about 9.5 and 28.8 μg m^{-3}, respectively) mean concentrations. The decrease in temperature range and wind speed (Figure 7) and the rise in relative humidity, alongside other meteorological parameters (MOFI, LIDS, and SHIF), indicate the stability of the atmosphere and cold weather-related conditions.

The third type of environment, leading to a decrease in B[a]P concentrations of 2.3 ng m^{-3}, was associated with medium mean PM$_{10}$, SO$_2$, and As levels (34.6 and 14.1 μg m^{-3}, and 1.6 ng m^{-3}, respectively), maximum observed B[a]P, As, and NO$_2$ concentrations (43.7 and 13.6 ng m^{-3}, and 98.5 μg m^{-3}, respectively), standard lifted index (304), and relative humidity (98%), minimum study period temperature (−15.3 °C), as well

as the highest number of precipitation events, i.e., non-zero TPP6, CPP6, and CRAI values (Figure 7).

The atmospheric stability and the intensity of anthropogenic emissions during the cold part of the year seem to result in high B[a]P concentrations. Since PAHs are mostly particle-bound, and precipitation scavenging plays a significant role in the PM removal from the atmosphere, it could be expected that wet deposition represents a way of PM-bound B[a]P elimination from the atmosphere. As shown by Liu et al. [76], wet removal and photodegradation are up to 10 and 5 times, respectively, more efficient in B[a]P elimination during summer than in winter. Additionally, wet scavenging dominates as a B[a]P removal path in summer, while the impact of photodegradation outweighs the wet removal in winter.

5.2. Arsenic

This study suggests that As concentrations affect B[a]P level dynamics up to 0.9 ng m^{-3} on average (Table 7), more than any other pollutant. A few types of environment were distinguished by analyzing the interrelations between As and B[a]P and their coexistence within certain conditions.

The obtained interrelation indicates similar emission sources of inorganic As, in a mixture of arsenite (AsIII) and arsenate (AsV), and organic B[a]P in the air, that could be identified as high-temperature combustion of fossil fuels and wood [77,78]. In addition, because of low volatility, both As and B[a]P mostly exist as particle-bound in the atmosphere, particularly associated with fine aerosol fractions. Up to approximately 10% of B[a]P occurs in the gaseous phase [79], although the multiphase B[a]P distribution was also highly dependent on ambient temperature [80].

In the first type of environment, B[a]P concentrations exhibited an increase in the range from 4 to 7 ng m^{-3} (Figure 8), with maximum concentrations reaching 30 ng m^{-3}. The relative impact of As, i.e., its association with B[a]P, compared to other studied parameters, reaches a maximum share of 43.6%. This environment was characterized by the lowest As and Cd concentrations, below 2 and 1.5 ng m^{-3}, respectively, and low to medium NOx, SO$_2$, and PM$_{10}$ levels of below 55 µg m^{-3}, below 25 µg m^{-3}, and from 10 to 25 µg m^{-3}, respectively. Other PM-bound constituents, including Ni and Pb, were registered in higher concentrations of 8.2 and 24.6 ng m^{-3}, respectively, which suggests the impact of local anthropogenic source emissions and dust resuspension, as well as the impact of occasional fossil fuel burning emissions. The co-occurrence of As and B[a]P was observed in the wide range of temperatures at the surface and 2 m (from 1 to 20 °C), which indicates that the relationship between As and B[a]P concentrations was not seasonally dependent. Additionally, this type of environment was featured by PBLH below 150 m, humidity above 74%, wind speed below 2 m s^{-1}, and very low MOFI (Figure 8), all of which reflected extremely stable meteorological and atmospheric conditions, which were registered on a few occasions during the measurement campaign. Therefore, it can be assumed that in the first type of environment, the contributions of remote air pollution sources and atmospheric long-range transport to the observed B[a]P and As concentrations can be excluded.
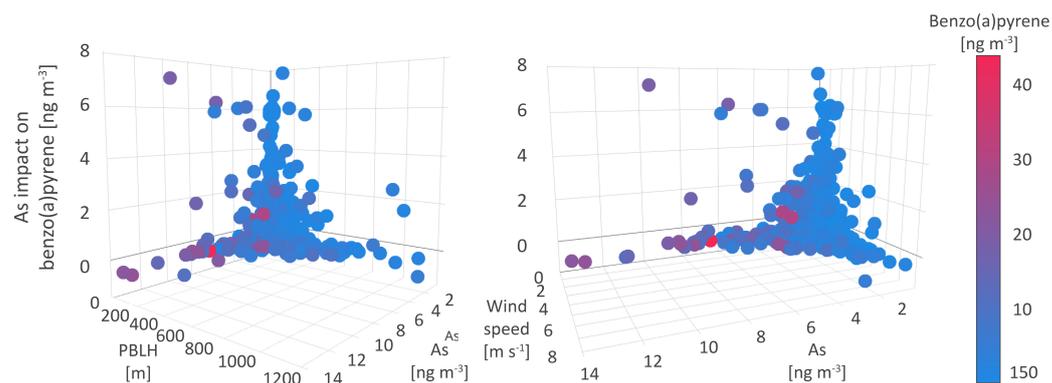
Figure 8. Arsenic impact on benzo(a)pyrene.

The second type of environment was characterized by an increase in B[a]P concentrations up to 4 ng m^{-3} on average and by the lower impact of As (5 to 20%), relative to other pollutants. In comparison to the previous one, this environment was also marked by up to three times higher PM$_{10}$ levels (70 µg m^{-3}), up to two times higher As (5 ng m^{-3}) and NOx (up to 100 µg m^{-3}) levels, and somewhat higher SO$_2$ (30 µg m^{-3}) concentrations. The assigned meteorological conditions included low humidity, air and soil temperatures ranging from -5 to 20 °C, PBLH below 480 m (Figure 8), and wind speed below 3.7 m s^{-1}, as well as MOFI values typical for the cold season. As can be concluded, the second type of environment represented the cold season and its associated emissions of As and B[a]P as well as inorganic oxides from heating-related sources. In cold weather conditions, PM, NOx, SO$_2$, and As are slow-reacting and the atmospheric reactions associated with the generation of secondary air pollutants (other oxide forms, sulfates, nitrates, or ozone), reaction byproducts or fine particles require a prolonged time, which in this case contributed to high pollutant concentrations assigned to the second type of B[a]P environment.

The third type of environment referring to the majority of measured pollutant concentrations recognized more than one pattern of As-B[a]P interrelations. Depending on the wind speed and other meteorological factors, both high and low B[a]P and As concentrations were registered. Namely, wind speed below 2 m s^{-1} was associated with the highest pollutant concentrations, while the increase in wind speed above 5 m s^{-1} resulted in a significant decrease in both pollutant concentrations below 1 ng m^{-3}. These findings suggest a negligible contribution of regional pollutant sources to air quality at the sampling site, but also the presence of local pollution sources and processes, such as resuspension of ash from crude-oil and lignite-fired boilers, which strongly affect pollutant concentrations during the episodes of low wind speed.

SHAP values ranging from -0.6 to 0 ng m^{-3} referred to the situations in which As levels had a moderately negative or null impact on B[a]P dynamics. On these occasions, As, B[a]P, and PM$_{10}$ levels were very high, 13.6 ng m^{-3}, 22 ng m^{-3} and 177 µg m^{-3}, respectively, while the SO$_2$ and NOx levels did not exceed 10 µg m^{-3}. Given these findings were associated with the T02M range from -3 to 5 °C, we can assume that As and B[a]P have separate sources during the cold season, which contribute to high concentrations of either one or another pollutant. More data and further analysis could provide detailed insight and confirm these assumptions.

5.3. Particulate Matter

The PM$_{10}$ concentration is the third significant parameter that affects B[a]P concentrations, as shown by the mean absolute SHAP value of 0.8 ng m^{-3}. In the absence of meteorological conditions favoring the association of B[a]P and small particle fraction, the relationship between PM$_{10}$ and B[a]P stands out.

The highest observed positive associations between PM$_{10}$ levels and B[a]P concentrations, in compliance with a relative share of 57.52% and assigned an absolute SHAP value of 8.36 ng m^{-3}, was registered in the environmental conditions associated with the lowest

concentrations of all pollutants, including PM_{10} levels below 32 μg m^{-3}. As regards meteorological conditions, the strongest interrelation between PM_{10} and B[a]P concentrations was detected in the environment characterized by air and soil temperatures ranging from 0 to 20 °C and low wind speed (below 2 m s^{-1}). This type of environment is not seasonally specific and might indicate natural interactions in the atmosphere such as associations between PAHs and PM. Atmospheric PAHs such as high-ring B[a]P are easily adsorbed onto suspended particles with high organic content [76] while the degradation of particle-bound B[a]P fraction is minimized or inhibited. The gas-to-particle partitioning of pollutants and atmospheric removal by wet scavenging are favored depending on the atmospheric conditions, PM surface, its composition and size, and contaminant properties [81]. In the warm season, the increase in temperatures leads to increased B[a]P volatility, followed by its biodegradation. As the impact of PM_{10} on B[a]P levels weaken the environmental conditions change slightly towards higher pollutant concentrations and an increase in wind speed and PBLH (Figure 9).
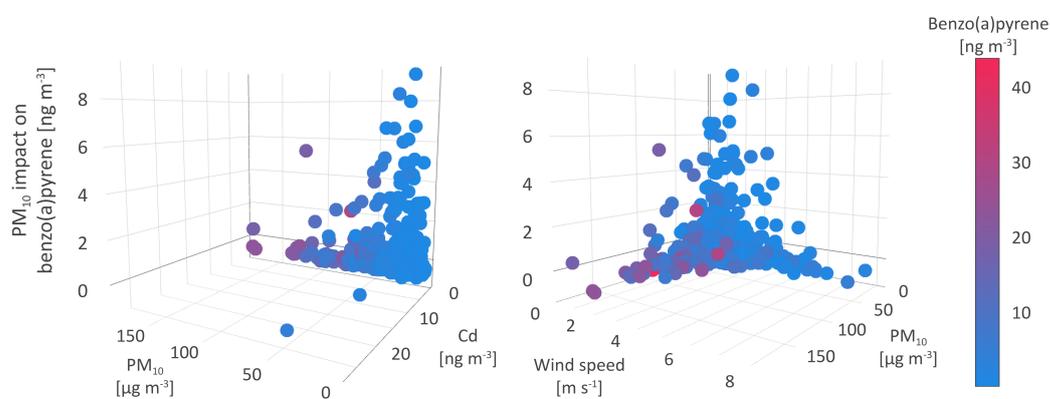

Figure 9. Particulate matter impact on benzo(a)pyrene.

Given the SHAP value of -0.87 PM_{10}, a high number of registered medium to high B[a]P concentrations was negatively associated with PM_{10}, particularly in the environment of high suspended particles As and low Cd, Ni, Pb, NOx, and SO$_2$ levels. As regards meteorological conditions, these interactions took place during the coldest days of the winter period, when low PBLH, high cloudiness, and wind speed up to 3 m s^{-1} were recorded (Figure 9).

As previously mentioned, the cold season was the period of intense emissions from power plants, domestic heating units, and commercial sources, resulting in elevated levels of PM, especially those of smaller diameter ($PM_{2.5}$ and PM_1) rather than PM_{10}. The finest particle fractions represented a highly suitable matrix for the adsorption of PAHs and these associations could be a possible explanation for the negative relation between B[a]P and PM. A number of studies have shown that small particle diameter plays an important role in the entrapment of PAHs, and thus more than 70% of high-molecular-weight PAHs with higher octanol-water partition coefficients, including B[a]P, is $PM_{2.5}$-bound [82,83]. Low air temperature, wind speed, solar radiation, and PBLH inhibited the vertical diffusion of pollutants and enhanced gas-to-particle pollutant partitioning [84,85]. In addition to this, the strong adsorption capacity of fine PM fraction prevailed over other environmental factors and suggests the particle partition of B[a]P to $PM_{2.5}$ and a smaller fraction rather than to PM_{10}. Lobscheid et al. [86] used multivariate linear regression models to predict relations of ambient B[a]P levels and $PM_{2.5}$ concentrations, spatial, temporal, and meteorological variates. The most significant variables included the average daily $PM_{2.5}$ concentration, wind speed, temperature, and relative humidity.

In contrast to this, during the warm and windy season, when the average temperatures, wind speed, and PBLH exceeded 15 °C, 4 m s^{-1}, and 450 m, respectively, the concentrations of PM_{10} and their constituents exhibited a significant decrease, although the same does not apply for NOx and SO$_2$. High solar radiation and temperature in warmer seasons lead to the

dispersion and photochemical degradation of the majority air polluting species [80,87], but the persistence of medium to high gaseous oxide levels during the warm season indicated the impact of intense and year-round continuous traffic emissions at the sampling site.

5.4. Nitrogen Oxides

Similar to PAHs, NOx (NO and smaller share of NO_2) emissions mainly resulted from the high-temperature combustion processes in power plants and motor vehicles. Both groups of compounds, PAHs and NOx, were subject to photochemical reactions in the atmosphere. Besides undergoing gas-particle phase distribution, PAHs are precursors for the generation of nitro-compounds. Namely, in the presence of free radicals, OH-PAH or NO_3-PAH are formed and subsequently, in the few-hour reaction with NO_2 upon release of nitric acid or water molecule, nitro-PAHs were generated [88,89].

The mean absolute SHAP value of 0.6 ng m^{-3} defines NOx as the third most significant parameter for shaping B[a]P levels in two distinguished types of environment, one of which strongly supports the increase in B[a]P concentrations. The polluted environment, with moderate to high B[a]P levels (average value of 3 ng m^{-3}) and attributed SHAP value of 6.78 ng m^{-3}, was characterized by a wide range of NOx, PM_{10}, and SO_2 concentrations, from 1.28 to 144 µg m^{-3}, up to 70 µg m^{-3} and up to 30 µg m^{-3}, respectively; however, the lowest levels of PM-bound As, Cd, Ni, and Pb (Figure 10).
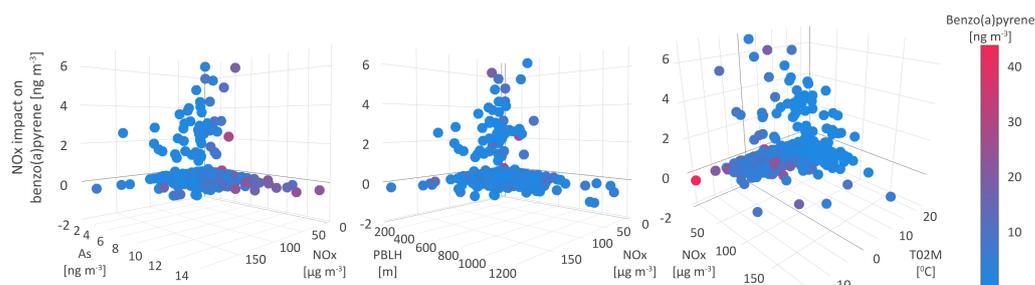

Figure 10. Nitrogen oxides impact on benzo(a)pyrene.

The meteorological conditions which enabled the positive impact of NOx on modelled B[a]P level dynamics and a wide range of B[a]P, NOx, PM_{10}, and SO_2 concentrations, refer to stable high-humidity cold weather without precipitations, with wind speed and PBLH below 3 m s^{-1} and 400 m, respectively; temperatures in the range from −7 to 20 °C, as well as with the corresponding CAPE, CPP6, CRAI, MOFI, and SHIF values (Figure 10). Under these conditions, common emission sources (fossil fuel burning for heating purposes) of the listed pollutants were intensified leading to their higher concentrations. In addition, the stagnant high-humidity conditions during heavy haze events enhanced the transformation of primary emitted particles containing PAHs to secondary organic aerosol (SOA), with the prominent presence of sulfate and nitrate water-soluble species dissolved in an aqueous outer particle layer [90].

The majority of studied pollutant events can be distinguished into two groups depending on the SHAP values and the strength of NOxs negative impact on modeled B[a]P concentration dynamics. The type of environment in which NOx and B[a]P interrelations are expressed by a lower SHAP value of −1.76 ng m^{-3}, refers to the warm season, with air temperatures from 15 to 20 °C, an occasional wind of high speed from 5 to 8 m s^{-1} and mean daily PBLH above 1000 m. As can be expected, these meteorological conditions have favored pollutant dispersion and resulted in low pollutant concentrations, as confirmed by measurements (Figure 10). During the warm season, PAHs undergo photolysis or processes which can yield their derivative compounds, such as oxygenated and nitrated PAHs. The UV-mediated ozone photolysis is a source of OH radicals in the troposphere, which react with PAHs to produce intermediate compounds OH-PAHs. After substitution with NO_2, OH-PAHs are further converted to nitro-PAHs, particularly at night, when the concentrations of NO are low [88]. Additionally, nitro-PAHs are also generated in the

chemical reactions between PAHs and NO_3-radicals, originating from reactions between O_3 and NO, and their formation can explain the negative NOx and B[a]P interrelations.

The SHAP value of -0.3 ng m^{-3} was attributed to the environment where no significant interactions between NOx and B[a]P were registered. Occasionally, these events were characterized either by high concentrations of B[a]P, PM$_{10}$, PM-bound constituents, and low NOx levels, or the opposite, the lowest concentrations of suspended particles, their constituents and high NOx levels exceeding 50 μg m^{-3}, which implies two different sources of origin.

6. Conclusions

In this study, we employed, coupled, and optimized advanced artificial intelligence-based modeling to accurately interrelate air pollution-related parameters to capture defining factors and processes that shape benzo(a)pyrene behavior. We have applied the XGBoost model optimized by metaheuristics and the Shapley Additive exPlanations explainable artificial intelligence method to a two-year database of pollutant concentrations and meteorological parameters to characterize types of environments that govern the interactions between benzo(a)pyrene, other polluting species, and meteorological conditions. The results suggest that the hybrid self-adaptive sine cosine algorithm method displayed a supreme performance level, by achieving the best scores for key performance indicators (mean square error of 2.5 and R^2 of 0.9), while the firefly algorithm scored the best results for standard deviation and variance, by delivering the most stable results. As shown, the temperature at the surface, arsenic, PM$_{10}$, and NOx were recognized to affect 22.7%, 14.4%, and 10.0% of benzo(a)pyrene concentrations, respectively. The observed interrelation between particulates and inorganic and organic pollutants could be associated with intensified fossil fuel burning such as low-quality lignite coal during the cold season. In the conditions of low temperature, PM, NOx, SO$_2$, and As are slow-reacting, and the atmospheric reactions in which the pollutants are involved require a prolonged time, which in this case enhanced the pollutant ambient levels. In addition, during cold seasons, photodegradation of B[a]P was weakened and its adsorption to the particles was favored. The results of this study have proved the potential of the applied methodology to improve the scientific knowledge and understanding of the complex factors that govern the environmental fate of air-polluting species.

Abbreviations

The following abbreviations are used in this manuscript:

SHAP	SHapley Additive exPlanations	
XGBoost	eXtreme Gradient Boosting	
Label	**Meteorological paramater**	**Unit**
PRSS	Pressure at surface	hPa
MSLP	Pressure reduced to mean sea level	hPa
TPP6	Accumulated precipitation (6 h accumulation)	m
MOFI	Momentum flux intensity (3- or 6-h average)	N m^{-2}
MOFD	Momentum flux direction (3- or 6-h average)	°

SHTF	Sensible heat net flux at surface (3- or 6-h average)	W m^{-2}
DSWF	Downward short wave radiation flux (3- or 6-h average)	W m^{-2}
RH2M	Relative Humidity at 2 m AGL	%
WS	Wind speed at 10 m AGL	m s^{-1}
WD	Wind direction at 10 m AGL	°
TO2M	Temperature at 2 m AGL	°C
TCLD	Total cloud cover (3- or 6-h average)	%
SHGT	Geopotential height	gpm *
CAPE	Convective available potential energy	J Kg^{-1}
CINH	Convective inhibition	J Kg^{-1}
LISD	Standard lifted index	°C
LIB4	Best 4-layer lifted index	°C
PBLH	Planetary boundary layer height	m
TMPS	Temperature at surface	°C
CPP6 **	Accumulated convective precipitation (6 h accumulation)	m
SOLM	Volumetric soil moisture content	frac.
CSNO	Categorial snow (yes = 1, no = 0) (3- or 6-h average)	
CICE	Categorial ice (yes = 1, no = 0) (3- or 6-h average)	
CFZR	Categorial freezing rain (yes = 1, no = 0) (3- or 6-h average)	
CRAI	Categorial rain (yes = 1, no = 0) (3- or 6-h average)	
LHTF	Latent heat net flux at surface (3- or 6-h average)	W/m^2
LCLD	Low cloud cover (3- or 6-h average)	%
MCLD	Middle cloud cover (3- or 6-h average)	%
HCLD	High cloud cover (3- or 6-h average)	%

* geopotential meters
** Beginning with 00 UTC July 15, 2019, CPPA (total accumulation) instead of CPP6
(6-h accumulation)

References

1. Ravina, M.; Esfandabadi, Z.S.; Panepinto, D.; Zanetti, M. Traffic-induced atmospheric pollution during the COVID-19 lockdown: Dispersion modeling based on traffic flow monitoring in Turin, Italy. *J. Clean. Prod.* **2021**, *317*, 128425. [CrossRef]
2. Hoffer, A.; Jancsek-Turóczi, B.; Tóth, Á.; Kiss, G.; Naghiu, A.; Levei, E.A.; Marmureanu, L.; Machon, A.; Gelencsér, A. Emission factors for PM 10 and polycyclic aromatic hydrocarbons (PAHs) from illegal burning of different types of municipal waste in households. *Atmos. Chem. Phys.* **2020**, *20*, 16135–16144. [CrossRef]
3. Mahasakpan, N.; Chaisongkaew, P.; Inerb, M.; Nim, N.; Phairuang, W.; Tekasakul, S.; Furuuchi, M.; Hata, M.; Kaosol, T.; Tekasakul, P.; et al. Fine and ultrafine particle-and gas-polycyclic aromatic hydrocarbons affecting southern Thailand air quality during transboundary haze and potential health effects. *J. Environ. Sci.* **2023**, *124*, 253–267. [CrossRef]
4. Stanišić, S.; Jovanović, G.; Perišić, M.; Snježana, H.R.; Milićević, T.; Stojić, A. Explaining the Environmental Fate of PAHs in Indoor and Outdoor Environments by the Use of Artificial Intelligence. In *Polycyclic Aromatic Hydrocarbons*; Gregoire, W.L., Ed.; Nova Science: Hauppauge, NY, USA, 2022; Chapter 1; pp. 1–36.
5. Karagulian, F.; Belis, C.A.; Dora, C.F.C.; Prüss-Ustün, A.M.; Bonjour, S.; Adair-Rohani, H.; Amann, M. Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level. *Atmos. Environ.* **2015**, *120*, 475–483. [CrossRef]
6. Elzein, A.; Stewart, G.J.; Swift, S.J.; Nelson, B.S.; Crilley, L.R.; Alam, M.S.; Reyes-Villegas, E.; Gadi, R.; Harrison, R.M.; Hamilton, J.F.; et al. A comparison of PM 2.5-bound polycyclic aromatic hydrocarbons in summer Beijing (China) and Delhi (India). *Atmos. Chem. Phys.* **2020**, *20*, 14303–14319. [CrossRef]
7. Marmett, B.; Carvalho, R.B.; Muccillo-Baisch, A.L.; Baisch, P.R.M.; Dos Santos, M.; Garcia, E.M.; Rhoden, C.R.; da Silva Júnior, F.M.R. Emissions monitoring and carcinogenic risk assessment of PM 10-bounded PAHs in the air from Candiota's coal activity area, Brazil. *Environ. Geochem. Health* **2023**, *45*, 899–911. [CrossRef] [PubMed]
8. Stojić, A.; Vuković, G.; Perišić, M.; Stanišić, S.; Šoštarić, A. Urban air pollution: An insight into its complex aspects. In *A Closer Look at Urban Areas*; Nova Science Publishers: Hauppauge, NY, USA, 2018.
9. Stojić, A.; Mustać, B.; Jovanović, G.; Đinović Stojanović, J.; Perišić, M.; Stanišić, S.; Herceg Romanić, S. Patterns of PCB-138 Bioaccumulation in Small Pelagic Fish from the Eastern Mediterranean Sea Using Explainable Machine Learning Prediction. In *Artificial Intelligence: Theory and Applications*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 175–189.

10. Stanišić, S.; Perišić, M.; Jovanović, G.; Maletić, D.; Vudragović, D.; Vranić, A.; Stojić, A. What Information on Volatile Organic Compounds Can Be Obtained from the Data of a Single Measurement Site Through the Use of Artificial Intelligence? In *Artificial Intelligence: Theory and Applications*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 207–225.

11. Stojić, A.; Jovanović, G.; Stanišić, S.; Romanić, S.H.; Šoštarić, A.; Udovičić, V.; Perišić, M.; Milićević, T. The PM2. 5-bound polycyclic aromatic hydrocarbon behavior in indoor and outdoor environments, part II: Explainable prediction of benzo [a] pyrene levels. *Chemosphere* **2022**, *289*, 133154. [CrossRef] [PubMed]

12. Šoštarić, A.; Stojić, S.S.; Vuković, G.; Mijić, Z.; Stojić, A.; Gržetić, I. Rainwater capacities for BTEX scavenging from ambient air. *Atmos. Environ.* **2017**, *168*, 46–54. [CrossRef]

13. Perišić, M.; Maletić, D.; Stojić, S.S.; Rajšić, S.; Stojić, A. Forecasting hourly particulate matter concentrations based on the advanced multivariate methods. *Int. J. Environ. Sci. Technol.* **2017**, *14*, 1047–1054. [CrossRef]

14. Stojić, A.; Maletić, D.; Stojić, S.S.; Mijić, Z.; Šoštarić, A. Forecasting of VOC emissions from traffic and industry using classification and regression multivariate methods. *Sci. Total. Environ.* **2015**, *521*, 19–26. [CrossRef]

15. Stojić, A.; Stanić, N.; Vuković, G.; Stanišić, S.; Perišić, M.; Šoštarić, A.; Lazić, L. Explainable extreme gradient boosting tree-based prediction of toluene, ethylbenzene and xylene wet deposition. *Sci. Total. Environ.* **2019**, *653*, 140–147. [CrossRef]

16. Mirjalili, S. SCA: A sine cosine algorithm for solving optimization problems. *Knowl.-Based Syst.* **2016**, *96*, 120–133. [CrossRef]

17. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.

18. Stegherr, H.; Heider, M.; Hähner, J. Classifying Metaheuristics: Towards a unified multi-level classification system. *Nat. Comput.* **2020**, *21*, 155–171. [CrossRef]

19. Emmerich, M.; Shir, O.M.; Wang, H. Evolution strategies. In *Handbook of Heuristics*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 89–119.

20. Fausto, F.; Reyna-Orta, A.; Cuevas, E.; Andrade, Á.G.; Perez-Cisneros, M. From ants to whales: Metaheuristics for all tastes. *Artif. Intell. Rev.* **2020**, *53*, 753–810. [CrossRef]

21. Beni, G. Swarm intelligence. In *Complex Social and Behavioral Systems: Game Theory and Agent-Based Models*; Springer: New York, NY, USA, 2020; pp. 791–818.

22. Abraham, A.; Guo, H.; Liu, H. Swarm intelligence: Foundations, perspectives and applications. In *Swarm Intelligent Systems*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 3–25.

23. Kennedy, J.; Eberhart, R. Particle swarm optimization. In Proceedings of the ICNN'95-International Conference on Neural Networks, Perth, WA, Australia, 27 November–1 December1995; Volume 4, pp. 1942–1948.

24. Dorigo, M.; Birattari, M.; Stutzle, T. Ant colony optimization. *IEEE Comput. Intell. Mag.* **2006**, *1*, 28–39. [CrossRef]

25. Yang, X.S. Firefly algorithms for multimodal optimization. In Proceedings of the International Symposium on Stochastic Algorithms, Sapporo, Japan, 26–28 October 2009; pp. 169–178.

26. Yang, X.S. A new metaheuristic bat-inspired algorithm. In *Nature Inspired Cooperative Strategies for Optimization (NICSO 2010)*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 65–74.

27. Yang, X.S.; Gandomi, A.H. Bat algorithm: A novel approach for global engineering optimization. *Eng. Comput.* **2012**, *29*, 464–483. [CrossRef]

28. Abualigah, L.; Diabat, A.; Mirjalili, S.; Abd Elaziz, M.; Gandomi, A.H. The arithmetic optimization algorithm. *Comput. Methods Appl. Mech. Eng.* **2021**, *376*, 113609. [CrossRef]

29. Wolpert, D.H.; Macready, W.G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1997**, *1*, 67–82. [CrossRef]

30. Zivkovic, M.; Bacanin, N.; Venkatachalam, K.; Nayyar, A.; Djordjevic, A.; Strumberger, I.; Al-Turjman, F. COVID-19 cases prediction by using hybrid machine learning and beetle antennae search approach. *Sustain. Cities Soc.* **2021**, *66*, 102669. [CrossRef] [PubMed]

31. Zivkovic, M.; Venkatachalam, K.; Bacanin, N.; Djordjevic, A.; Antonijevic, M.; Strumberger, I.; Rashid, T.A. Hybrid Genetic Algorithm and Machine Learning Method for COVID-19 Cases Prediction. In *Proceedings of the International Conference on Sustainable Expert Systems: ICSES 2020*; Springer Nature: Singapore, 2021; Volume 176, p. 169.

32. Bacanin, N.; Bezdan, T.; Tuba, E.; Strumberger, I.; Tuba, M.; Zivkovic, M. Task scheduling in cloud computing environment by grey wolf optimizer. In Proceedings of the 2019 27th Telecommunications Forum (TELFOR), Belgrade, Serbia, 26–27 November 2019; pp. 1–4.

33. Bezdan, T.; Zivkovic, M.; Tuba, E.; Strumberger, I.; Bacanin, N.; Tuba, M. Multi-objective Task Scheduling in Cloud Computing Environment by Hybridized Bat Algorithm. In Proceedings of the International Conference on Intelligent and Fuzzy Systems, Istanbul, Turkey, 21–23 July 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 718–725.

34. Bezdan, T.; Zivkovic, M.; Antonijevic, M.; Zivkovic, T.; Bacanin, N. Enhanced Flower Pollination Algorithm for Task Scheduling in Cloud Computing Environment. In *Machine Learning for Predictive Analysis*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 163–171.

35. Zivkovic, M.; Bezdan, T.; Strumberger, I.; Bacanin, N.; Venkatachalam, K. Improved Harris Hawks Optimization Algorithm for Workflow Scheduling Challenge in Cloud–Edge Environment. In *Computer Networks, Big Data and IoT*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 87–102.

36. Zivkovic, M.; Bacanin, N.; Tuba, E.; Strumberger, I.; Bezdan, T.; Tuba, M. Wireless Sensor Networks Life Time Optimization Based on the Improved Firefly Algorithm. In Proceedings of the 2020 International Wireless Communications and Mobile Computing (IWCMC), Limassol, Cyprus, 15–19 June 2020; pp. 1176–1181.

37. Zivkovic, M.; Bacanin, N.; Zivkovic, T.; Strumberger, I.; Tuba, E.; Tuba, M. Enhanced Grey Wolf Algorithm for Energy Efficient Wireless Sensor Networks. In Proceedings of the 2020 Zooming Innovation in Consumer Technologies Conference (ZINC), Novi Sad, Serbia, 26–27 May 2020; pp. 87–92.

38. Bacanin, N.; Tuba, E.; Zivkovic, M.; Strumberger, I.; Tuba, M. Whale Optimization Algorithm with Exploratory Move for Wireless Sensor Networks Localization. In Proceedings of the International Conference on Hybrid Intelligent Systems, Bhopal, India, 10–12 December 2019; pp. 328–338.

39. Zivkovic, M.; Zivkovic, T.; Venkatachalam, K.; Bacanin, N. Enhanced Dragonfly Algorithm Adapted for Wireless Sensor Network Lifetime Optimization. In *Data Intelligence and Cognitive Informatics*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 803–817.

40. Bezdan, T.; Cvetnic, D.; Gajic, L.; Zivkovic, M.; Strumberger, I.; Bacanin, N. Feature Selection by Firefly Algorithm with Improved Initialization Strategy. In Proceedings of the 7th Conference on the Engineering of Computer Based Systems, Novi Sad Serbia, 26–27 May 2021; pp. 1–8.

41. Bezdan, T.; Zivkovic, M.; Tuba, E.; Strumberger, I.; Bacanin, N.; Tuba, M. Glioma Brain Tumor Grade Classification from MRI Using Convolutional Neural Networks Designed by Modified FA. In Proceedings of the International Conference on Intelligent and Fuzzy Systems, Istanbul, Turkey, 21–23 July 2020; pp. 955–963.

42. Zivkovic, M.; Bacanin, N.; Antonijevic, M.; Nikolic, B.; Kvascev, G.; Marjanovic, M.; Savanovic, N. Hybrid CNN and XGBoost Model Tuned by Modified Arithmetic Optimization Algorithm for COVID-19 Early Diagnostics from X-ray Images. *Electronics* **2022**, *11*, 3798. [CrossRef]

43. Strumberger, I.; Tuba, E.; Zivkovic, M.; Bacanin, N.; Beko, M.; Tuba, M. Dynamic search tree growth algorithm for global optimization. In *Proceedings of the Doctoral Conference on Computing, Electrical and Industrial Systems*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 143–153.

44. Preuss, M.; Stoean, C.; Stoean, R. Niching Foundations: Basin Identification on Fixed-Property Generated Landscapes. In Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation, Dublin, Ireland, 12–16 July 2011; pp. 837–844. [CrossRef]

45. Jovanovic, D.; Antonijevic, M.; Stankovic, M.; Zivkovic, M.; Tanaskovic, M.; Bacanin, N. Tuning Machine Learning Models Using a Group Search Firefly Algorithm for Credit Card Fraud Detection. *Mathematics* **2022**, *10*, 2272. [CrossRef]

46. Petrovic, A.; Bacanin, N.; Zivkovic, M.; Marjanovic, M.; Antonijevic, M.; Strumberger, I. The AdaBoost Approach Tuned by Firefly Metaheuristics for Fraud Detection. In Proceedings of the 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), Sonbhadra, India, 17–19 June 2022; pp. 834–839.

47. Bacanin, N.; Sarac, M.; Budimirovic, N.; Zivkovic, M.; AlZubi, A.A.; Bashir, A.K. Smart wireless health care system using graph LSTM pollution prediction and dragonfly node localization. *Sustain. Comput. Inform. Syst.* **2022**, *35*, 100711. [CrossRef]

48. Jovanovic, L.; Jovanovic, G.; Perisic, M.; Alimpic, F.; Stanisic, S.; Bacanin, N.; Zivkovic, M.; Stojic, A. The Explainable Potential of Coupling Metaheuristics-Optimized-XGBoost and SHAP in Revealing VOCs' Environmental Fate. *Atmosphere* **2023**, *14*, 109. [CrossRef]

49. Bacanin, N.; Zivkovic, M.; Stoean, C.; Antonijevic, M.; Janicijevic, S.; Sarac, M.; Strumberger, I. Application of Natural Language Processing and Machine Learning Boosted with Swarm Intelligence for Spam Email Filtering. *Mathematics* **2022**, *10*, 4173. [CrossRef]

50. Stankovic, M.; Antonijevic, M.; Bacanin, N.; Zivkovic, M.; Tanaskovic, M.; Jovanovic, D. Feature Selection by Hybrid Artificial Bee Colony Algorithm for Intrusion Detection. In Proceedings of the 2022 International Conference on Edge Computing and Applications (ICECAA), Tamilnadu, India, 13–15 October 2022; pp. 500–505.

51. Milosevic, S.; Bezdan, T.; Zivkovic, M.; Bacanin, N.; Strumberger, I.; Tuba, M. Feed-Forward Neural Network Training by Hybrid Bat Algorithm. In Proceedings of the Modelling and Development of Intelligent Systems: 7th International Conference, MDIS 2020, Sibiu, Romania, 22–24 October 2020; Revised Selected Papers 7; Springer International Publishing: New York, NY, USA, 2021; pp. 52–66.

52. Gajic, L.; Cvetnic, D.; Zivkovic, M.; Bezdan, T.; Bacanin, N.; Milosevic, S. Multi-layer Perceptron Training Using Hybridized Bat Algorithm. In *Computational Vision and Bio-Inspired Computing*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 689–705.

53. Bacanin, N.; Zivkovic, M.; Al-Turjman, F.; Venkatachalam, K.; Trojovskỳ, P.; Strumberger, I.; Bezdan, T. Hybridized sine cosine algorithm with convolutional neural networks dropout regularization application. *Sci. Rep.* **2022**, *12*, 6302. [CrossRef]

54. Bacanin, N.; Stoean, C.; Zivkovic, M.; Jovanovic, D.; Antonijevic, M.; Mladenovic, D. Multi-Swarm Algorithm for Extreme Learning Machine Optimization. *Sensors* **2022**, *22*, 4204. [CrossRef] [PubMed]

55. Jovanovic, L.; Jovanovic, D.; Bacanin, N.; Jovancai Stakic, A.; Antonijevic, M.; Magd, H.; Thirumalaisamy, R.; Zivkovic, M. Multi-Step Crude Oil Price Prediction Based on LSTM Approach Tuned by Salp Swarm Algorithm with Disputation Operator. *Sustainability* **2022**, *14*, 14616. [CrossRef]

56. Bukumira, M.; Antonijevic, M.; Jovanovic, D.; Zivkovic, M.; Mladenovic, D.; Kunjadic, G. Carrot grading system using computer vision feature parameters and a cascaded graph convolutional neural network. *J. Electron. Imaging* **2022**, *31*, 061815. [CrossRef]

57. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017.

58. Molnar, C. *Interpretable Machine Learning*; Lulu. com: Morrisville, NC, USA, 2020.

59. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef]

60. Derrac, J.; García, S.; Molina, D.; Herrera, F. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm Evol. Comput.* **2011**, *1*, 3–18. .: 10.1016/j.swevo.2011.02.002. [CrossRef]

61. de Mattos Neto, P.S.; Marinho, M.H.; Siqueira, H.; de Souza Tadano, Y.; Machado, V.; Antonini Alves, T.; de Oliveira, J.F.L.; Madeiro, F. A methodology to increase the accuracy of particulate matter predictors based on time decomposition. *Sustainability* **2020**, *12*, 7310. [CrossRef]

62. Neto, P.S.D.M.; Firmino, P.R.A.; Siqueira, H.; Tadano, Y.D.S.; Alves, T.A.; De Oliveira, J.F.L.; Marinho, M.H.D.N.; Madeiro, F. Neural-based ensembles for particulate matter forecasting. *IEEE Access* **2021**, *9*, 14470–14490. [CrossRef]

63. Goldberg, D.E.; Richardson, J. Genetic algorithms with sharing for multimodal function optimization. In *Genetic algorithms and their applications: Proceedings of the Second International Conference on Genetic Algorithms*; Lawrence Erlbaum: Hillsdale, NJ, USA, 1987; Volume 4149.

64. Mirjalili, S. Genetic algorithm. In *Evolutionary Algorithms and Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 43–55.

65. Karaboga, D.; Basturk, B. On the performance of artificial bee colony (ABC) algorithm. *Appl. Soft Comput.* **2008**, *8*, 687–697. [CrossRef]

66. Mirjalili, S.; Lewis, A. The whale optimization algorithm. *Adv. Eng. Softw.* **2016**, *95*, 51–67. [CrossRef]

67. Heidari, A.A.; Faris, H.; Aljarah, I.; Mirjalili, S.; Mafarja, M.; Chen, H. Harris hawks optimization: Algorithm and applications. *Future Gener. Comput. Syst.* **2019**, *97*, 849–872. [CrossRef]

68. Khishe, M.; Mosavi, M.R. Chimp optimization algorithm. *Expert Syst. Appl.* **2020**, *149*, 113338. [CrossRef]

69. LaTorre, A.; Molina, D.; Osaba, E.; Poyatos, J.; Del Ser, J.; Herrera, F. A prescription of methodological guidelines for comparing bio-inspired optimization algorithms. *Swarm Evol. Comput.* **2021**, *67*, 100973. [CrossRef]

70. Glass, G.V. Testing homogeneity of variances. *Am. Educ. Res. J.* **1966**, *3*, 187–190. [CrossRef]

71. Shapiro, S.S.; Francia, R. An approximate analysis of variance test for normality. *J. Am. Stat. Assoc.* **1972**, *67*, 215–216. [CrossRef]

72. Hsu, H.; Lachenbruch, P.A. Paired t test. In *Wiley StatsRef: Statistics Reference Online*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2014.

73. Chen, H.; Yang, B.; Wang, S.j.; Wang, G.; Liu, D.y.; Li, H.; Liu, W.b. Towards an optimal support vector machine classifier using a parallel particle swarm optimization strategy. *Appl. Math. Comput.* **2014**, *239*, 180–197. [CrossRef]

74. Zhao, Y.; Wang, L.; Luo, J.; Huang, T.; Tao, S.; Liu, J.; Yu, Y.; Huang, Y.; Liu, X.; Ma, J. Deep learning prediction of polycyclic aromatic hydrocarbons in the high arctic. *Environ. Sci. Technol.* **2019**, *53*, 13238–13245. [CrossRef]

75. Cao, X.; Huo, S.; Zhang, H.; Ma, C.; Zheng, J.; Wu, F.; Song, S. Seasonal variability in multimedia transport and fate of benzo [a] pyrene (BaP) affected by climatic factors. *Environ. Pollut.* **2022**, *292*, 118404. [CrossRef]

76. Liu, C.; Guo, Y.; Shi, K.; Zhang, J.; Wu, B.; Du, J. Comparative analysis of contributions of wet deposition and photodegradation to the removal of atmospheric BaP by MFDCCA. *Sci. Rep.* **2021**, *11*, 5515. [CrossRef]

77. Moreno, N.; Viana, M.; Pandolfi, M.; Alastuey, A.; Querol, X.; Chinchon, S.; Pinto, J.F.; Torres, F.; Díez, J.M.; Saez, J. Determination of direct and fugitive PM emissions in a Mediterranean harbour by means of classic and novel tracer methods. *J. Environ. Manag.* **2009**, *91*, 133–141. [CrossRef] [PubMed]

78. Millán-Martínez, M.; Sánchez-Rodas, D.; de la Campa, A.S.; Alastuey, A.; Querol, X.; Jesús, D. Source contribution and origin of PM10 and arsenic in a complex industrial region (Huelva, SW Spain). *Environ. Pollut.* **2021**, *274*, 116268. [CrossRef]

79. Guerreiro, C.; Horálek, J.; de Leeuw, F.; Couvidat, F. Benzo (a) pyrene in Europe: Ambient air concentrations, population exposure and health effects. *Environ. Pollut.* **2016**, *214*, 657–667. [CrossRef] [PubMed]

80. Hunová, I.; Kurfürst, P.; Vlasáková, L.; Schreiberová, M.; Škáchová, H. Atmospheric Deposition of Benzo [a] pyrene: Developing a Spatial Pattern at a National Scale. *Atmosphere* **2022**, *13*, 712. [CrossRef]

81. Hu, T.; Mao, Y.; Ke, Y.; Liu, W.; Cheng, C.; Shi, M.; Zhang, Z.; Zhang, J.; Qi, S.; Xing, X. Spatial and seasonal variations of PAHs in soil, air, and atmospheric bulk deposition along the plain to mountain transect in Hubei province, central China: Air-soil exchange and long-range atmospheric transport. *Environ. Pollut.* **2021**, *291*, 118139. [CrossRef]

82. Jakovljević, I.; Pehnec, G.; Vađić, V.; Čačković, M.; Tomašić, V.; Jelinić, J.D. Polycyclic aromatic hydrocarbons in PM 10, PM 2.5 and PM 1 particle fractions in an urban area. *Air Qual. Atmos. Health* **2018**, *11*, 843–854. [CrossRef]

83. Jia, J.; Deng, L.; Bi, C.; Jin, X.; Zeng, Y.; Chen, Z. Seasonal variations, gas-PM2. 5 partitioning and long-distance input of PM2. 5-bound and gas-phase polycyclic aromatic hydrocarbons in Shanghai, China. *Atmos. Environ.* **2021**, *252*, 118335. [CrossRef]

84. Lv, Y.; Li, X.; Xu, T.T.; Cheng, T.T.; Yang, X.; Chen, J.M.; Iinuma, Y.; Herrmann, H. Size distributions of polycyclic aromatic hydrocarbons in urban atmosphere: Sorption mechanism and source contributions to respiratory deposition. *Atmos. Chem. Phys.* **2016**, *16*, 2971–2983. [CrossRef]

85. Nguyen, T.N.T.; Jung, K.S.; Son, J.M.; Kwon, H.O.; Choi, S.D. Seasonal variation, phase distribution, and source identification of atmospheric polycyclic aromatic hydrocarbons at a semi-rural site in Ulsan, South Korea. *Environ. Pollut.* **2018**, *236*, 529–539. [CrossRef] [PubMed]
86. Lobscheid, A.B.; McKone, T.E.; Vallero, D.A. Exploring relationships between outdoor air particulate-associated polycyclic aromatic hydrocarbon and PM2. 5: A case study of benzo (a) pyrene in California metropolitan regions. *Atmos. Environ.* **2007**, *41*, 5659–5672. [CrossRef]
87. Siudek, P. Polycyclic aromatic hydrocarbons in coarse particles (PM10) over the coastal urban region in Poland: Distribution, source analysis and human health risk implications. *Chemosphere* **2023**, *311*, 137130. [CrossRef]
88. Lee, Y.Y.; Hsieh, Y.H.; Huang, B.W.; Mutuku, J.K.; Chang-Chien, G.P.; Huang, S. An Overview: PAH and Nitro-PAH Emission from the Stationary Sources and their Transformations in the Atmosphere. *Aerosol Air Qual. Res.* **2022**, *22*, 220164. [CrossRef]
89. Lammel, G.; Mulder, M.D.; Shahpoury, P.; Kukučka, P.; Lišková, H.; Přibylová, P.; Prokeš, R.; Wotawa, G. Nitro-polycyclic aromatic hydrocarbons–gas–particle partitioning, mass size distribution, and formation along transport in marine and continental background air. *Atmos. Chem. Phys.* **2017**, *17*, 6257–6270. [CrossRef]
90. Wang, J.; Ye, J.; Zhang, Q.; Zhao, J.; Wu, Y.; Li, J.; Liu, D.; Li, W.; Zhang, Y.; Wu, C.; et al. Aqueous production of secondary organic aerosol from fossil-fuel emissions in winter Beijing haze. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2022179118. [CrossRef]