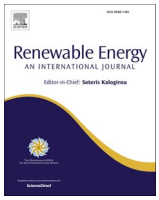
Renewable Energy 232 (2024) 121066

Contents lists available at ScienceDirect

Renewable Energy

journal homepage: www.elsevier.com/locate/renene



Examining effects of air pollution on photovoltaic systems via interpretable random forest model

Adam Dudáš [a], Mihaela Tinca Udristioiu [b,*], Tarik Alkharusi [c,e], Hasan Yildizhan [c,d], Satheesh Kumar Sampath [e]

[a] Department of Computer Science, Faculty of Natural Sciences, Matej Bel University, Banska Bystrica, Slovakia
[b] Department of Physics, Faculty of Physics, University of Craiova, Craiova, Romania
[c] Clean Energy Processes (CEP) Laboratory, Department of Chemical Engineering, Imperial College London, London, UK
[d] Department of Energy Systems Engineering, Adana Alparslan Türkeş Science and Technology University, Adana, Türkiye
[e] Engineering Department, University of Technology and Applied Sciences, Muscat, Oman

ARTICLE INFO

Keywords:
Air pollution
Particulate matter
Interpretable machine
Photovoltaic systems

ABSTRACT

Renewable energy plays a vital role in power generation and solar photovoltaic systems due to resource availability throughout the year. This work aims to investigate the impact of air pollutants and meteorological parameters on the performance of the photovoltaic systems locally, taking into consideration the advantages of the photovoltaic power potential of the SW part of Romania, where Craiova is located (average solar radiation intensity >1350 kWh/m^2/year). This study is based on a one-year dataset provided by a sensor that monitors particulate matter concentrations, volatile organic compounds, dioxide of carbon, ozone, noise, formaldehyde and three climate parameters (temperature, pressure, and relative humidity). The research methodology applies an innovative interpretable random forest model emphasising the implications of air pollution for photovoltaic systems. The proposed machine learning model was trained to predict the particulate matter level in air based on the basic environmental variable measurements. The study presents six random forest models of varying complexity, which reach the accuracy of classification for the selected problem up to 99 %, and applies the Shapley Additive Explanations technique to interpret the decision-making model. The observation regarding the highest concentration of particulate matter occurring during cold months, which typically do not align with peak solar irradiance, underscores the importance of considering various environmental factors in solar energy planning. With its practical implications, this insight offers decision-makers valuable information about the feasibility of optimising solar energy generation despite seasonal variations in air pollution levels, directly addressing their needs and concerns.

Nomenclature

Abbreviation	
WHO	World Health Organization
EEA	European Environment Agency
Symbols	
PM	Particulate matter [μg/m^3]
PM1	Particulate matter of a diameter of 1 μm
PM2.5	Particulate matter of a diameter of 2.5 μm
PM10	Particulate matter of a diameter of 10 μm
T	Temperature [°C]
P	Pressure [Pa]
RH	Relative Humidity [%]
O_3	Ozone [DU]
CO_2	Carbon Dioxide [ppm]
CH_2O	Formaldehyde [ppm]
VOC	Volatile Organic Compounds [μg/m^3]
E_{th}	The amount of energy generated [kWh]
E_{glob}	Golabl solar radiation [W/m^2]
A_{array}	photovoltaic cell area [m^2]
Greek symbol	
η	photovoltaic cell efficiency [W/m^2]
r	Pearson correlation coefficient
ρ	Spearman rank correlation coefficient
ϕ	Shapley value

(*continued on next column*)

(*continued*)

* Corresponding author.
E-mail address: mtudristioiu@central.ucv.ro (M.T. Udristioiu).

https://doi.org/10.1016/j.renene.2024.121066
Received 16 May 2024; Received in revised form 18 July 2024; Accepted 22 July 2024
Available online 23 July 2024

1. Introduction

In 2021, the World Health Organization (WHO) significantly decreased the limits of pollutants for the protection of health to 15 μg/m^3, which is the average value for PM2.5 concentrations in 24 h. Also, 45 μg/m^3 is the average value for PM10 concentration in 24 h, having a maximum of 3–4 exceedance days per year with no recommendation for PM1 concentration [1]. The European Environment Agency (EEA) set lower pollutant concentration thresholds than the WHO but established that the authorities should implement an air quality management plan in case of exceedances. The one-year average for PM2.5 concentrations is 25 μg/m^3, and for PM10, the average concentration is recommended to be 50 μg/m^3 in 24 h; for one year, the average is recommended to be 40 μg/m^3. These values should not exceed 35 days/per year [2].

Air pollution poses a severe environmental threat to human health; continuous exposure to PM2.5 has more health effects than PM10. PM1 needs to be regulated and well-studied. The study conducted in Beijing revealed that PM1 might have a more significant negative impact on health than exposure to PM2.5 and PM10 [3,4]. The link between short-term PM1 exposure and cardio-respiratory diseases has been reported [5]. Long-term PM1 exposure affects lung function in children and adolescents more than in the short term [6]. The short-term effects of PM1 and PM2.5 air pollution were investigated based on hospital admission for respiratory diseases [7–9].

One of the solutions to reduce air pollution is to replace conventional energy sources such as coal, oil, and gas with electrical power plants. Transitioning from conventional to renewable power plants, such as solar power plants, could significantly mitigate air pollution. Solar power is a clean, sustainable energy option that produces electricity without emitting harmful pollutants or greenhouse gases. By harnessing the abundant energy from the sun, we can reduce our reliance on fossil fuels and decrease air pollution levels. Implementing solar power for public institutions and transportation systems is a great step towards a cleaner and more sustainable future. Solar panels can be installed on rooftops, parking lots, and other suitable locations to generate electricity locally, reducing the need for energy from polluting power plants.

Additionally, integrating solar energy into transportation infrastructure, such as electric buses or solar-powered charging stations for electric vehicles, can further reduce emissions from conventional transportation methods. Also, these power plants have some limitations: (1) solar power technology depends on the weather and extreme climatic conditions [10,11]; (2) there is seasonal climate variability [12]; (3) there are geographic limitations to the development of solar plants [13]; (4) air pollution affects solar power generation by reducing the amount of solar energy that reaches the photovoltaic surface via reflection, scattering, and absorption [14–16].

The enhanced interpretability of data-driven air quality models might be utilised to identify pollution sources and quantify source contributions [17]. In [18], the key drivers influencing air pollution were analysed using interpretable algorithms. A Spatial-Temporal Interpretable Deep Learning Model for ground-level PM2.5 retrievals using Moderate Resolution Imaging Spectroradiometer imagery was developed to train the model with data for three years and apply it in the fourth year of data [19]. In [20], a random forest model was developed to estimate monthly the PM2.5 concentrations at 1 km resolution and covering a large area in North China. Gap-filled AOD, MERRA-2 simulations, meteorological parameters, and land cover as predictors are included to obtain a high prediction accuracy of 0.88 (R^2).

In [21,22], four machine learning models are applied on an installed 30.6 kWh/day and identified redundant energy on the mini-grid in the 56.98–119.86 kWh/day range. Further analysis shows that redundant energy can support the demand for household cooking energy through sustainable thermal batteries. Among four machine learning models, K-Nearest Neighbors Regression is at the top. Most machine learning approaches in air pollution and solar energy-related ways focus on predicting specific values (the regression task) [23]. Here, the first research gap is identified; instead of regression, the classification task is used as the first novelty element of this paper. The following section discusses the influence of particulate matter on the amount of energy produced using photovoltaic systems.

The presence of PM2.5 directly affects electricity generation in photovoltaic systems. **Equation (1)** can be used to estimate the amount of energy generated (E_{th}) in kWh.

$$E_{th} = \eta E_{glob} A_{array} \tag{1}$$

where η is photovoltaic cell efficiency, E_{glob} is solar radiation measured in W/m^2, and A_{array} is photovoltaic cell area measured in m^2 [24].

High particulate matter concentrations significantly influence incoming solar irradiation by absorbing and/or scattering the sunlight, especially in heavily polluted areas [25]. Therefore, the amount of electrical energy generated by a photovoltaic system is not certain; the efficiency of this type of system can be influenced by the number of environmental variables, such as the amount of solar irradiation to the earth, temperature, humidity, dust, and finally, particulate matter levels [24,26,27]. According to a study developed in a 600 kW solar photovoltaic power plant, the authors noticed that PV modules that have been unclean for over a year might lose up to 5.66 % of their power output [28]. Snow removal from PV arrays helps to avoid PV damage and enhance power output, while panel cleaning improves energy efficiency. Besides relying on natural cleaning processes (wind, rain, and melting snow), alternative solutions like manual cleaning, installing PV with hydrophobic and hydrophilic coatings, and electrodynamic screens reduce particulate matter (PM) concentration during the cold months, enhancing the performance of PV systems [29]. Increasing the inclination angle relative to the horizontal surface minimises the requirement to remove snow from the solar panel [30].

Table 1 shows the amount of solar irradiation lost because of PM2.5 levels. Air pollution significantly affects the energy performance of photovoltaic systems. Even at low levels of pollution (PM2.5 ≈ 12 μg/m^3→level 1), the energy performance of the systems decreases by approximately 2.5 %. With pollution that can be classified as level 2, an energy performance loss of around 5 % was recorded. From this literature, a second research gap is identified and filled with a machine learning-based approach to particulate matter level estimation using interpretable decision trees and random forest classification models.

The summary of the contributions is as follows.

1. Statistical analysis and pruning of sizable environmental datasets useable in any machine learning model. The dataset comprises one year's worth of sensor measurements focused on several environmental variables that could be significant from the point of view of photovoltaic systems. The measurements were taken by a sensor placed in the center of Craiova, Romania.
2. A machine learning model is built based on an interpretable decision tree/random forest method, which uses a created dataset to classify the input into one of the considered particulate matter levels (Table 1). This model is evaluated based on its classification accuracy, precision, and interpretability of the decisions made.

Table 1
Particulate matter levels and their effect on solar irradiation [25].

Level of air pollution	PM2.5 ($\mu g/m^3$)	PM10 ($\mu g/m^3$)	Solar irradiation effect (average value)
1 - good	0–30	0–50	≈ 2.5 % loss
2 - satisfactory	31–60	51–100	≈ 5 % loss
3 - moderate	61–90	101–250	≈ 8 % loss
4 - poor	91–120	251–350	≈ 12.5 % loss
5 - very poor	121–250	351–430	≈ 25 % loss
6 - severe	251+	431+	+ ≈ 12.5 % per each 100 μg/m^3

3. The model shows the impact of air pollutants and meteorological parameters on the performance of the photovoltaic systems at the local level, considering the advantages of the photovoltaic potential of the area where Craiova is located.

Besides the introduction and conclusion, the article has two main sections. **Section 2** describes the statistical evaluation and process of the environmental variable dataset collected using a low-cost sensor located in the center of Craiova City, Romania. **Section 3** first part focuses on describing the classification system containing a random forest model and Shapley value interpretation module. In the second part of this section, an evaluation of the proposed model is presented with the performance metrics of accuracy, precision, and interpretability of the system's decision-making process.

2. Environmental variable dataset

2.1. Sensor and measurements dataset description

The dataset was provided by a monitoring system model uRADMonitor A3 (made by Magnasci SRL, Romania), which belongs to the University of Craiova. The monitoring system is part of the network of sensors *Clear Air Oltenia*, built in the framework of academic-community cooperation [23,31,32]. Data measurements were taken every minute from December 5, 2021, to December 6, 2022. Model uRADMonitor A3 (ID 820002C3, latitude 44.3194, longitude 23.8011, altitude 120 m) is an automated, fixed air pollution monitoring system located in the center of Craiova, on the exterior wall of the mentioned university, straight to a busy traffic intersection and near an underground passage. The dataset used in this study was collected by the monitoring system with the identity number ID 820002C3 in the first two years after its installation. The automated system measured three climate parameters such as temperature (T), pressure (P), Relative Humidity (RH), three particulate matter concentrations (PM1, PM2.5, PM10), ozone (O_3), carbon dioxide (CO_2), formaldehyde (CH_20), volatile organic compounds (VOC), and noise. PM concentrations are determined using the laser scattering method, which differs from the gravimetric method used by the official stations (based on the weight difference of filters pre- and post-sampling). Carbon dioxide concentration is measured by a non-dispersive infrared sensor. Two electrochemical sensors track formaldehyde and ozone. A metal-oxide sensor determines volatile organic compounds, and the noise level is measured using an analogic sensor. The dataset includes the local time for all measurements. (Refer to Appendix A for more technical specifications about the sensor model uRADMonitor A3).

Two independent international laboratories (National Research and Development Institute for Industrial Ecology, Romania, and Air Quality Sensor Performance Evaluation Center, USA) tested the accuracy of the sensor model uRADMonitor A3, and the results are made public on the manufacturer's site. The manufacturer calibrated the sensor before selling it, based on a comparison with the reference sensor (the corrections were included in the equipment's software, as the two labs recommended).

2.2. Statistical analysis and pruning of the dataset

The raw dataset has 458 916 records, each containing a value measured for the following properties: date and time of measurement, T, P, RH, PM1, PM2.5, PM10, CO_2, O_3, VOC, noise, and CH_2O.

Data preprocessing has been done to remove the outliers in the dataset, and Fig. 1 presents such outlying values for the PM2.5 and PM10 attributes. It is observed that almost all the measurements are present in the interval of values between 0 and 200 $\mu g/m^3$. However, there are occasional measurements that are present entirely outside of the body of the dataset. Since the number of such measurements was not high (281 specifically), it was decided to prune the dataset to remove these outliers. This pruning process resulted in a dataset of 458 635 records containing no outliers, disseminated in Fig. 2.

Also, in preprocessing, the dataset has been aligned to perform the specific task: the classification of data based on particulate matter levels such as PM2.5 and PM10 attributes, as specified in Table 1. The data is classified into one of the six defined pollution classes using the IF-THEN rules. By observing the data, it contained five levels of PM2.5 pollution, which means the sensor measured good to very poor levels. Also, for PM10, there are only three levels: from good to moderate. The statistical description of the pruned dataset is presented in Table 2, which consists of minimal (min) and maximal (max) values of an attribute, the median and mean of an attribute, and the first and third quartile of an attribute value interval. A correlation analysis on the pruned dataset is performed to measure the predictive potential needed to build machine learning models. So, the Pearson correlation coefficient and Spearman rank correlation coefficient are considered relevant representatives of correlation coefficients.

(i) ***Pearson correlation coefficient*** focuses on the linear prediction of values. For the relationship between attributes A and B, it is described using the following **Equation (2)** [33]:

$$r = \frac{\sum_{i=1}^{n}(A_i - \mu(A))(B_i - \mu(B))}{\sqrt{\sum_{i=1}^{n}(A_i - \mu(A))^2}\sqrt{\sum_{i=1}^{n}(B_i - \mu(B))^2}} \tag{2}$$

where $\mu(A)$ is the mean of attribute A, similarly $\mu(B)$ is the mean value of attribute B, and n is the number of measurements (vertical size of the dataset). This apparent dependence on the mean value brings the most significant disadvantage of the Pearson correlation coefficient - sensitivity to outliers.

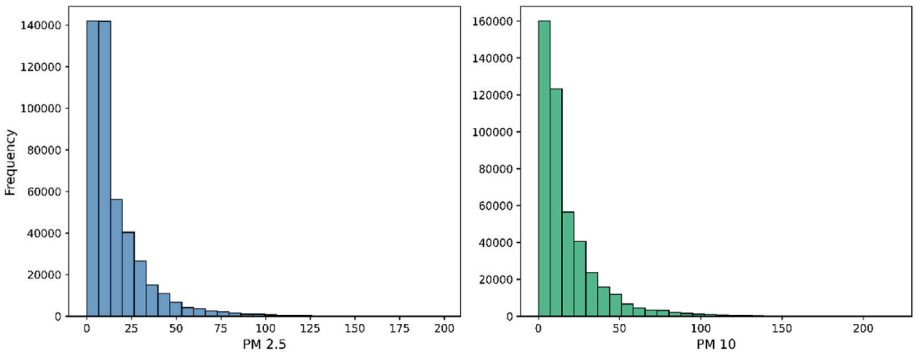


Fig. 1. PM2.5 (left) and PM10 (right) frequency in the raw dataset.

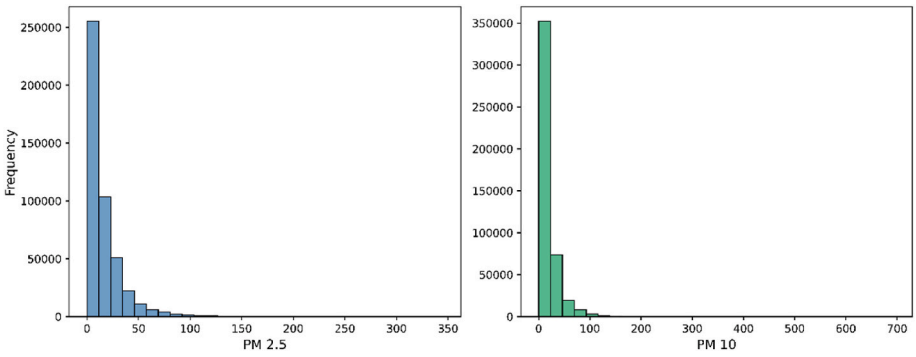


Fig. 2. PM2.5 (left) and PM10 (right) frequency in the pruned dataset.

Table 2
Summary statistics of the pruned dataset.

	min	1st quart	median	mean	3rd quart	max
temperature	−9.43	4.25	11.4	12.16	19.37	36.34
pressure	98 114	99 932	100 389	100 430	100 933	103 000
humidity	51.9	70.4	78.4	77.25	84.4	98.9
CO_2	400	528	587	584	641	891
O_3	20	20	20	25.43	28	89
VOC	30 761	191 593	217 422	213 836	239 112	319 755
noise	27.35	54.85	57.35	57.33	59.85	88.85
CH_2O	10	13	14	14.12	15	19
$class_{2.5}$	1	1	–	–	1	5
$class_{10}$	1	1	–	–	1	3

(ii) ***Spearman rank correlation coefficient*** measures the monotonicity of the values within the attribute. This type of correlation coefficient is not recommended if repeated values in the dataset effect are attenuated with increasing dataset size. Spearman rank correlation coefficient is computed as [33]:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{3}$$

where $d = rank\,(a_i)\; rank\,(b_i)$ and n is the number of considered attributes.

Figs. 3 and 4 present a correlation heatmap for Pearson and Spearman rank correlation coefficients measured in both PM2.5 and PM10 datasets. It is observed that the correlation between environmental variables and the classification of individual measurements (last column or row of the matrix) is of interest. In both PM2.5 and PM10, the correlation coefficient values are close to the extremes of the considered interval (1 or −1), marked by the darkest places in the correlation heatmap. The most significant correlations between environmental variables and classification are the following.

a) Pearson correlation measured between VOC and classification of measurement for PM2.5 dataset, where $r = -0.512$
b) Spearman rank correlation measured between temperature and classification of measurement for PM2.5 dataset, where $\rho = -0.328$
c) Pearson correlation measured between VOC and classification of measurement for PM10 dataset, where $r = -0.449$
d) Spearman rank correlation measured between temperature and classification of measurement for PM10 dataset, where $\rho = -0.228$

These low correlations between variables of interest point to the need for complex classifiers, which are not dependent on linear functions, average values, or the normal distribution of data in the considered space.

3. Interpretable decision tree and random forest model for particulate matter classification

Interpretable classifiers are used to estimate the level of air pollution in the considered area. Fig. 5 describes the proposed machine learning system, which can be divided into two modules as follows.

1. ***Data processing module***, which works with raw sensor data and uses the process described in Section 2.2 to prune the data with the objective of classification of the data. After this process, the pruned dataset consists of 10 attributes used in the next module.

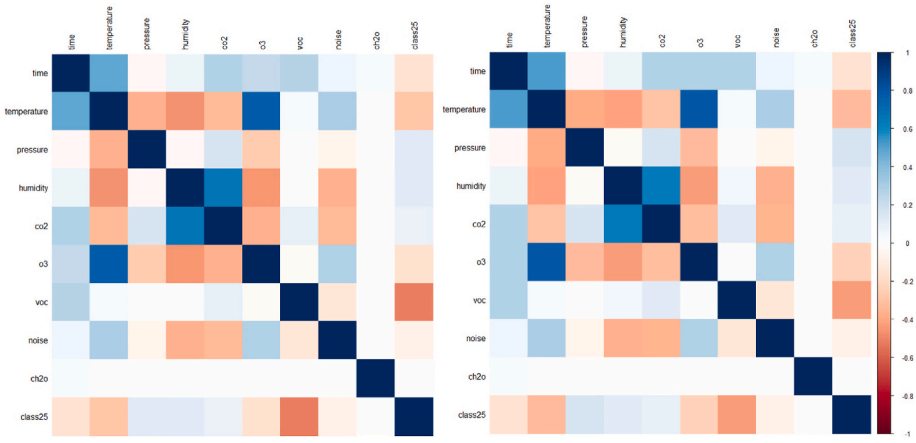


Fig. 3. Comparison of Pearson (left) and Spearman Rank (right) correlation matrices for the PM2.5 version of the dataset.

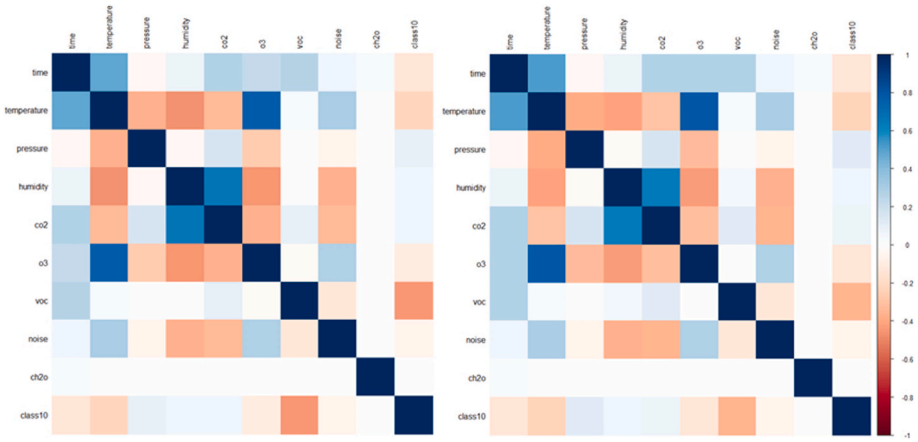


Fig. 4. Comparison of Pearson (left) and Spearman Rank (right) correlation matrices for the PM10 version of the dataset.

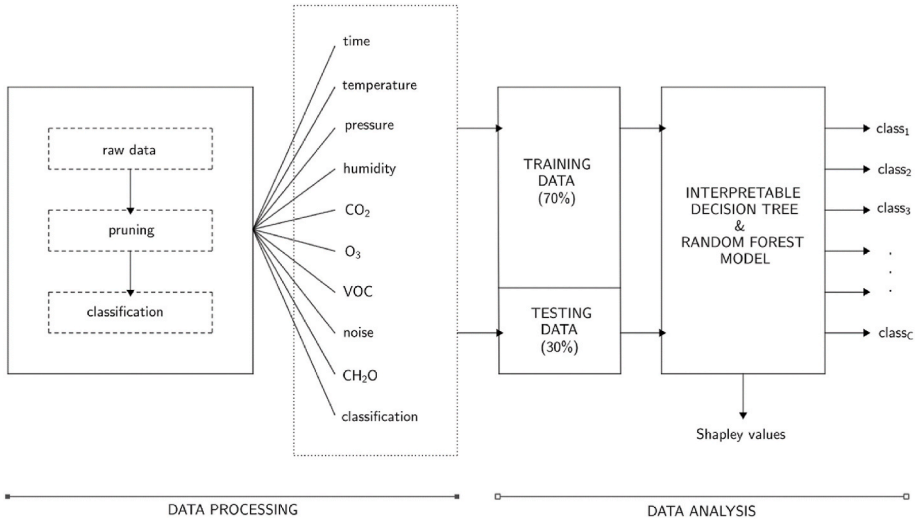


Fig. 5. Schema of the proposed system for particulate matter level classification.

2. ***Data analysis module*** uses a standard sequence of tasks to classify the data into one of the proposed classes. More specifically:
   - The dataset created in the data processing module is randomly divided into two data subsets - training data, on which the random forest model is learned, and testing data, which are used to evaluate the quality of the predictions of the created model. The ratio of the distribution of input data to the training and testing subset is 70 % to 30 %.
   - Both subsets of the input data are subsequently used in the classification module of the system. This module is based on a random forest classifier consisting of various numbers of decision trees. The output of this module is the classification of the input graph into one of the considered classes of air pollution level.
   - Since the random forest model achieves a high classification accuracy, but the interpretability of its decisions is very complex, the proposed model includes an interpretation module. This interpretation of the decision-making process is performed based on the analysis of the contribution of individual graph properties to the final classification result and the evaluation done through Shapley values.

3.1. Evaluation of quality of random forest model

Each dataset's classification accuracy and precision are calculated to evaluate the decision-making quality of the created random forest model. By accuracy, the decision-making quality of the model is the closeness of the predicted value to the real value of the feature. Precision refers to the ability of the created classification model to identify only the relevant entities [34].

Confusion matrices for all measurements can compute the accuracy and precision of the random forest model. Accuracy is computed based on the confusion matrix for each dataset as given in **Equation (4)**.

$$accuracy_S^n = \frac{t_n + t_p}{t_n + t_p + f_p + f_n} \tag{4}$$

where S is the size of particulate matter, n is the size of the forest, t_n is the number of true negative samples, t_p is the number of true positive samples, f_p is the number of falsely positive samples, and f_n is the number of false negative samples. The precision of the created model is computed for each class individually by **Equation (5)**.

$$precision_S^n(C) = \frac{t_p}{t_p + t_n} \tag{5}$$

where S is the size of particulate matter, n is the size of the forest, C is considered class (level of particulate matter pollution), t_p is the number of true positive samples, and t_n is the number of true negative samples.

3.2. Evaluation of interpretability of random forest model

The Shapley Additive Explanations technique uses Shapley values to interpret the decisions of the created model. This method measures how individual environmental variables (features) or sets of features contribute to the overall quality of the created random forest model. The Shapley value ϕ for the i-th environmental variable is computed in Equation (6) [35–37].

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|(M-|z'|-1)!}{M!} [f_x(z') - f_x(z' \backslash i)] \quad (6)$$

where x is the measurement considered by the model, f is the created random forest model, z' is the subset of features whose contribution to the decision is measured, x' is the simplified form of the measurement x (sum of its' features) and M is the number of features active in the used model.

With Shapley values, the average contribution of an environmental variable (or feature) value is measured to the prediction in various combinations of features. Shapley value distribution was also measured for environmental variables of both PM2.5 and PM10 classifications.

4. Results and discussion

4.1. Performance assessment of particulate matter 2.5 for different random forest size

Confusion matrices for the created model working with pollution levels of PM2.5 are given in Tables 3–8. These tables contain a few correctly and incorrectly classified records and differ in the number of trees in the forest.

Table 3 contains the confusion matrix of a single decision tree for PM2.5 level classification. As can be seen from the accuracy of 1.5 computed from this matrix, the system reaches high accuracy even with the single decision tree. However, the quality of classification to individual classes - precision - is somewhat lacking; in some classes, it is as low as 76 %. Table 4 contains the confusion matrix for random forest consisting of two decision trees for PM2.5 level classification. The model's accuracy is similar to the previous case, but a more precise classification of individual classes can be noticed.

Table 5 describes the confusion matrix for the random forest model, which contains five decision trees. This model size is the first to reach 99 % classification accuracy, but some classes with less-than-acceptable precision, namely classes 3 and 4, are the most problematic.

The confusion matrix in Table 6 describes the performance of a random forest composed of ten trees. Curiously, the accuracy of the forest decreased to 98 %. Similarly, this model's average precision is also lower than in the previous case. In the case of 50 tree random forest, Table 7 shows the strongest accuracy and precision of classification yet. Since the size of the random forest is not high, continue with more trees.

Table 8 presents the classification quality of the last experiments. An increase in accuracy or precision was observed. In this case, a random forest contains 100 trees and reaches the same results as the previous model with one slight increase in precision in classifying records to class number 4. After this experiment, the accuracy and precision values either stagnated or decreased with the increased size of the forest.

For PM2.5, the accuracy of the classification model consisting of a single decision tree is 97 %, and precision for classes 1 to 5 is 99 %, 85 %, 79 %, 76 % and 80 %, respectively. By increasing the decision tree of two, the accuracy is 97 %, and precision for classes 1 to 5 is 98 %, 91 %, 87 %, 83 % and 96 %, respectively. If the forest size is increased to five, the accuracy is 99 %, and precision for classes 1 to 5 is 99 %, 92 %, 90 %, 88 % and 93 %, respectively. Still, attempts have been made for ten decision trees, and the accuracy is 98 % and precision for classes 1 to 5 is 99 %, 95 %, 91 %, 89 % and 96 %, respectively.

For the same PM2.5, the accuracy of the classification model for fifty decision trees is 99 %, and the precision for classes 1 to 5 is 99 %, 95 %, 93 %, 91 % and 95 %, respectively. By increasing the decision tree to one hundred, the accuracy is 99 %, and the precision for classes 1 to 5 is 99

Table 3
Confusion matrix of decision tree for PM2.5 level classification.

	$class_1$	$class_2$	$class_3$	$class_4$	$class_5$
$\boldsymbol{class_1}$	117 622	1624	58	1	10
$\boldsymbol{class_2}$	1569	11 987	436	23	16
$\boldsymbol{class_3}$	35	400	2382	134	13
$\boldsymbol{class_4}$	10	36	144	640	47
$\boldsymbol{class_5}$	1	4	10	41	348

Table 4
Confusion matrix of random forest model for PM2.5 level classification, size of the forest = 2.

	$class_1$	$class_2$	$class_3$	$class_4$	$class_5$
$\boldsymbol{class_1}$	118 871	440	3	1	0
$\boldsymbol{class_2}$	2505	11 411	111	4	0
$\boldsymbol{class_3}$	97	581	2240	46	0
$\boldsymbol{class_4}$	17	43	201	605	11
$\boldsymbol{class_5}$	0	11	30	73	290

Table 5
Confusion matrix of random forest model for PM2.5 level classification, size of the forest = 5.

	$class_1$	$class_2$	$class_3$	$class_4$	$class_5$
$\boldsymbol{class_1}$	118 567	743	5	0	0
$\boldsymbol{class_2}$	1115	12 738	175	2	1
$\boldsymbol{class_3}$	19	289	2596	58	2
$\boldsymbol{class_4}$	3	9	105	736	24
$\boldsymbol{class_5}$	0	2	2	36	364

Table 6
Confusion matrix of random forest model for PM2.5 level classification, size of the forest = 10.

	$class_1$	$class_2$	$class_3$	$class_4$	$class_5$
$\boldsymbol{class_1}$	118 813	494	8	0	0
$\boldsymbol{class_2}$	1054	12 819	157	1	0
$\boldsymbol{class_3}$	5	244	2663	52	0
$\boldsymbol{class_4}$	2	3	96	759	17
$\boldsymbol{class_5}$	0	1	1	38	364

Table 7
Confusion matrix of random forest model for PM2.5 level classification, size of the forest = 50.

	$class_1$	$class_2$	$class_3$	$class_4$	$class_5$
$\boldsymbol{class_1}$	118 802	508	3	2	0
$\boldsymbol{class_2}$	839	13 056	135	0	1
$\boldsymbol{class_3}$	2	218	2690	52	2
$\boldsymbol{class_4}$	0	5	79	774	19
$\boldsymbol{class_5}$	0	0	0	25	379

Table 8
Confusion matrix of random forest model for PM2.5 level classification, size of the forest = 100.

	$class_1$	$class_2$	$class_3$	$class_4$	$class_5$
$\boldsymbol{class_1}$	118 792	520	2	1	0
$\boldsymbol{class_2}$	828	13 064	137	2	0
$\boldsymbol{class_3}$	3	205	2709	45	2
$\boldsymbol{class_4}$	0	2	68	789	18
$\boldsymbol{class_5}$	0	0	0	24	380

%, 95 %, 93 %, 92 % and 95 %, respectively.

Fig. 6 presents the accuracy and precision values for the size of the random forest. As can be seen, all the measured values did not behave monotonously, mainly in the interval of 1–10 trees - but overall, with the growing size of the forest, accuracy and precision increased.

4.2. Performance assessment of particulate matter 10 for different random forest size

Confusion matrices for the created model working with pollution levels of PM10 are given in Tables 9–14. As mentioned, this problem is slightly different from the previous classification since the sensor identified only three levels of PM10 pollution (compared to five levels of PM2.5 pollution). As in the previous case, these tables contain several correctly and incorrectly classified records and differ in the number of trees in the forest. By observation, values of the confusion matrix presented in Table 9 describe classification using a single decision tree. As can be seen, the accuracy is very high even with this simple system, but precision is lacking in two of three classes. Table 10 contains values of the classification of data with the use of a random forest containing two decision trees. This model is significantly more precise in the act of classification compared to the single decision tree.

The model composed of five tree random forests presented in Table 11 is the first one to reach 100 % precision of classification of data instances to one of the considered classes (class 1) and a decrease in the precision of classification in the case of class 3. In the case of 10 tree random forest, Table 12 shows strong accuracy and precision of classification. Since there is still only 94 and 96 % classification precision for classes 2 and 3, we increase the random forest size. Table 13 contains a confusion matrix for 50 tree random forest and the same accuracy of the model but an increase in the precision of the classification in class 2. The last experiment presented in Table 14 was done on a 100-tree random forest, which reached the same classification results as the previous (50-tree) model.

For PM10, the accuracy of the classification model consisting of a single decision tree is 99 %, and the precision for classes 1 to 3 is 99 %, 86 % and 86 %, respectively. By increasing the decision tree of two, the accuracy is 99 %, and precision for classes 1 to 3 is 99 %, 92 % and 96 %, respectively. If the forest size is increased to five, the accuracy is 99 %, and precision for classes 1 to 3 is 100 %, 93 % and 94 %, respectively. Still, an attempt has been made for ten decision trees, where the accuracy is 99 % and the precision for classes 1 to 3 is 100 %, 94 % and 96 %, respectively. For fifty decision trees, the accuracy is 99 %, and precision for classes 1 to 3 is 100 %, 95 % and 96 %, respectively.

If forest size is increased to one hundred, the accuracy is 99 %, and precision for classes 1 to 3 is 100 %, 95 % and 96, respectively. The classification of the problem is summarised in Fig. 7, and the classification accuracy is equal to 99 % from the first experiment, which classified data using a single decision tree. However, for the precision to be as good as possible, we used more complex systems - more precisely, 2,

Table 9
Confusion matrix of decision tree for PM10 level classification.

	$class_1$	$class_2$	$class_3$
$\boldsymbol{class_1}$	128 288	836	27
$\boldsymbol{class_2}$	776	6136	170
$\boldsymbol{class_3}$	10	160	1188

Table 10
Confusion matrix of random forest model for PM10 level classification, size of the forest = 2.

	$class_1$	$class_2$	$class_3$
$\boldsymbol{class_1}$	128 926	224	1
$\boldsymbol{class_2}$	1240	5799	43
$\boldsymbol{class_3}$	31	292	1035

Table 11
Confusion matrix of random forest model for PM10 level classification, size of the forest = 5.

	$class_1$	$class_2$	$class_3$
$\boldsymbol{class_1}$	128 802	343	6
$\boldsymbol{class_2}$	497	6516	69
$\boldsymbol{class_3}$	4	115	1239

Table 12
Confusion matrix of random forest model for PM10 level classification, size of the forest = 10.

	$class_1$	$class_2$	$class_3$
$\boldsymbol{class_1}$	128 879	266	6
$\boldsymbol{class_2}$	554	6478	50
$\boldsymbol{class_3}$	3	131	1224

Table 13
Confusion matrix of random forest model for PM10 level classification, size of the forest = 50.

	$class_1$	$class_2$	$class_3$
$\boldsymbol{class_1}$	128 914	234	3
$\boldsymbol{class_2}$	429	6601	52
$\boldsymbol{class_3}$	3	95	1260

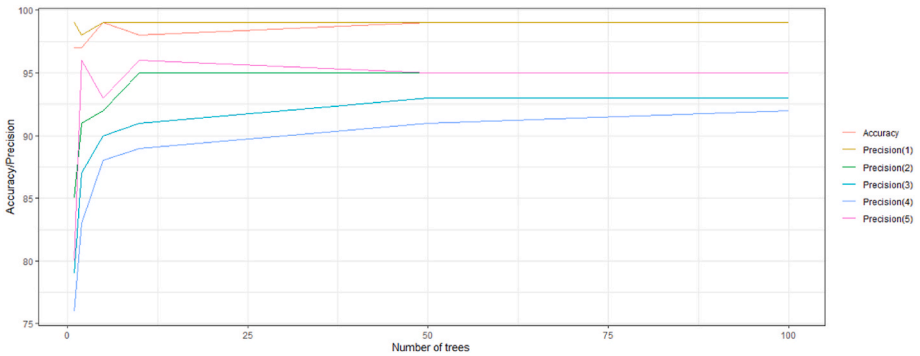


Fig. 6. Comparison of accuracy and precision of PM2.5 level classification with the use of 1, 2, 5, 10, 50 and 100 trees in random forest.

Table 14
Confusion matrix of random forest model for PM10 level classification, size of the forest = 100.

	$class_1$	$class_2$	$class_3$
$\boldsymbol{class_1}$	128 919	229	3
$\boldsymbol{class_2}$	400	6634	48
$\boldsymbol{class_3}$	1	85	1272

5, 10, 50, and 100 trees in a random forest. The best results were achieved using a random forest of the size 50 (or 100).

4.3. Analysis of most influential parameters for different sizes of particulate matter

The five influenced Shapley values for PM2.5 level classification are given in Fig. 8, and the most important are presented as follows.

(i) VOC - high values of this environmental variable bring slightly negative Shapley values, while mean values of VOC reach Shapley values close to 0. Low values of the feature are scattered on the positive side of the Shapley value spectrum.
(ii) Time - Shapley values of measured time bring some ambiguity to the model - low values of this environmental variable are present on the positive and negative sides of the Shapley values spectrum, with mean time values clustered in weak negative Shapley values. High values of the property are scattered mainly on the right side of the interval.
(iii) Temperature: The Shapley values for temperature are represented in the model as follows: high-temperature values reach negative Shapley values, low values of this environmental variable influence the classification ability of the model positively, and mean values are scattered throughout the Shapley value spectrum.
(iv) Pressure: High-pressure values are present on both sides of the Shapley value interval, with the mean and low values of the feature clustered close to 0 with some outliers.
(v) Relative Humidity: The model uses values of measured relative humidity like pressure, with lower humidity values being more prevalent on the slightly negative side of Shapley values.

The five strongest Shapley values for PM10 level classification are given in Fig. 9 and distributed similarly to the previous case. One difference is that the fifth most influential environmental variable for the created model was CO_2 with highly ambiguous behaviour: a mix of positive and negative Shapley values with high, mean, and low CO_2 values. A number of these values was clustered close to the Shapley value of 0.

Unsupervised machine learning algorithms were applied to analyse spatiotemporal air pollution patterns [38]. Authors analysed PM10 type of pollution in a large dataset collected from almost 100 sensors located in Krakow, over one year, with data being recorded at 1-h intervals – even though the presented study measured similar data in the same amount of time, the sizes of the Craiova dataset and Krakow dataset are comparable due to measurement interval in Craiova dataset being 1 min.

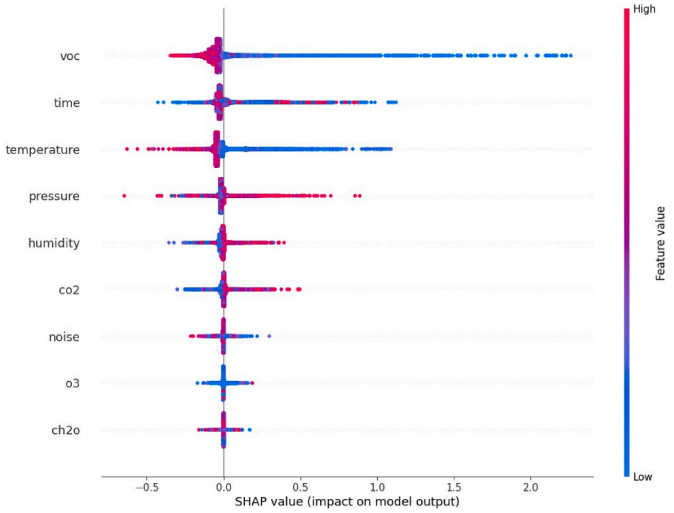


Fig. 8. Visualizations of Shapley values for individual features contributing to decision-making in the classification of PM2.5 pollution.

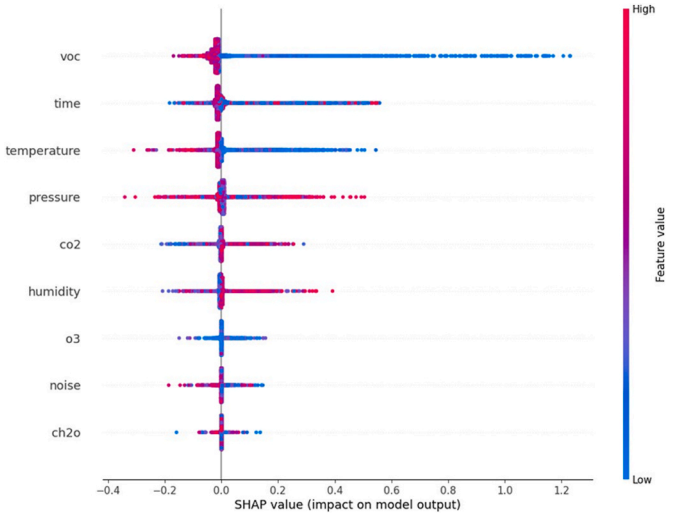


Fig. 9. Visualisation of Shapley values for individual features contributing to decision-making in the classification of PM10 pollution.

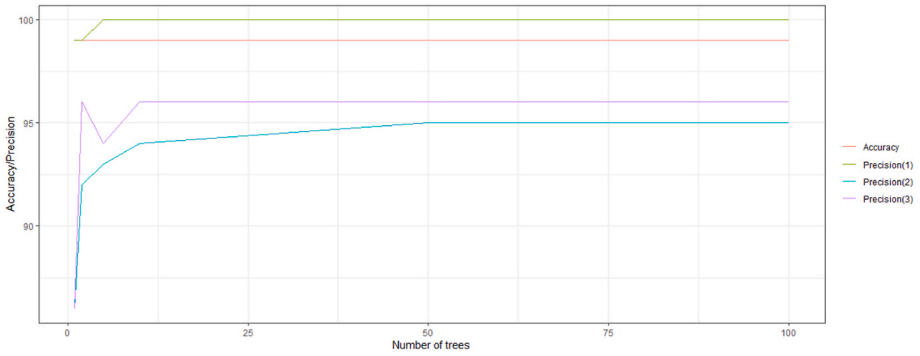


Fig. 7. Comparison of accuracy and precision of PM10 level classification using 1, 2, 5, 10, 50 and 100 trees in random forest.

The potential of machine learning techniques and big data analysis is highlighted for identifying air pollution patterns and informing policy decisions related to urban planning, traffic management, and public health interventions [38].

Based on the predicted results, it is observed that higher particulate matter concentrations occur during the cold months, which do not align with the peak times for solar irradiance, as illustrated in Fig. 10. This demonstrates that the utilisation of solar energy during peak periods remains unaffected by the concentration of particulate matter, which we subjectively categorise as level 1 (good), resulting in minimal losses during these peak times. However, during the winter season, when particulate matter concentrations are elevated, it is possible to suggest that natural cleaning processes, such as rain or wind, may reduce the deposition of such particulate matter on solar system surfaces. This reduction in deposition could potentially enhance the amount of light reaching these systems.

5. Conclusions

Innovative interpretable decision trees and random forest classification models are used to estimate air pollution levels. PM2.5 and PM10 pollution are classified with high accuracy and precision; also, the scope of this study is disseminated as follows.

1. For PM2.5, an accuracy of 99 % and precision of 99, 95, 93, 92, and 95 % for pollution levels 1 to 5, respectively, are obtained. The model used to reach these results consisted of a 100-tree random forest classifier (Fig. 6).
2. For PM10 level estimation, the created 50-tree random forest model reached values of accuracy equal to 99 % and precision of 100 % for level 1 pollution, 95 % for level 2 pollution, and 96 % for level 3 pollution (Fig. 7).
3. An interpretation of the classification with the Shapley value model for identifying crucial environmental variables, as seen in Figs. 8 and 9. Besides the classification model, we also created a pruned dataset containing one year's worth of environmental measurements, which can be used in computer-oriented experiments.
4. During the peak times of solar irradiance, it is observed that the PM concentration is at a minimum, which supports the utilisation of solar energy systems (Fig. 10).
5. Results of this study show that dynamic models using time-lagged data outperform static and reduced machine learning models. Incorporating time-lagged data in some cases improved the accuracy of machine learning models 3-fold compared to static and reduced models.

In the long term, these results might inspire decision-makers to develop solar photovoltaic parks as part of the transition to a more sustainable energy system and simultaneously solve the air pollution issue.

CRediT authorship contribution statement

Adam Dudáš: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Mihaela Tinca Udristioiu:** Writing – original draft, Supervision, Investigation, Data curation, Conceptualization. **Tarik Alkharusi:** Writing – review & editing, Visualization, Investigation, Conceptualization. **Hasan Yildizhan:** Writing – review & editing, Visualization, Conceptualization, Supervision. **Satheesh Kumar Sampath:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

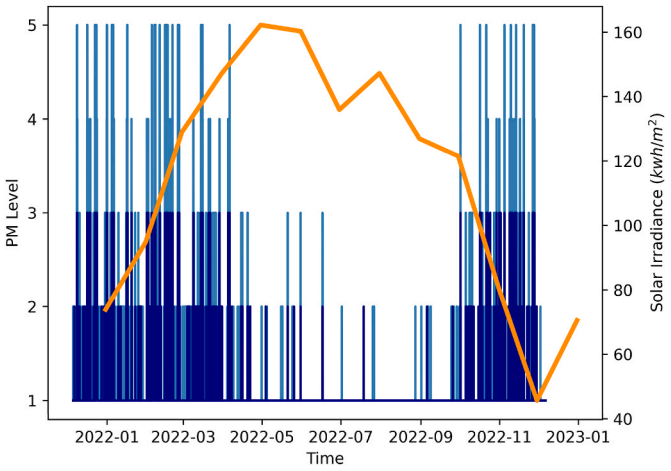


Fig. 10. Comparison of PM levels for both PM 2.5 (light blue) and PM 10 (dark blue) with monthly average Solar irradiance (orange).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.renene.2024.121066.

References

[1] WHO Ambient Air Quality Database Update, World Health Organization, 2021 n.d. URL, https://cdn.who.int/media/docs/default-source/air-pollution-documents/air-quality-and-health/who-air-quality-database-2021_final.pdf?sfvrsn=2371d57c_5. (Accessed 7 April 2023).
[2] Air quality standards - European Environment Agency. [www document], n.d. URL. https://www.eea.europa.eu/themes/air/air-quality-concentrations/air-quality-standards (accessed 7/April/2023).
[3] H. Wang, F. Lu, M. Guo, W. Fan, W. Ji, Z. Dong, Associations between PM1 exposure and daily emergency department visits in 19 hospitals, Beijing, Sci. Total Environ. 755 (1) (2021) 142507, https://doi.org/10.1016/j.scitotenv.2020.142507.
[4] L. Wang, D. Luo, X. Liu, J. Zhu, F. Wang, B. Li, L. Li, Effects of PM2.5 exposure on reproductive system and its mechanisms, Chemosphere 264 (1) (2021) 128436, https://doi.org/10.1016/j.chemosphere.2020.128436.
[5] L. Mei, S. Yan, Y. Li, X. Jin, X. Sun, Y. Wu, Y. Liang, Q. Wei, W. Yi, R. Pan, Yangyang He, Tang Chao, X. Liu, J. Cheng, H. Su, Q. Xu, Association between short-term PM1 exposure and cardiorespiratory diseases: evidence from a systematic review and meta-analysis, Atmos. Pollut. Res. 13 (1) (2022) 101254, https://doi.org/10.1016/j.apr.2021.101254.
[6] Z. Zong, M. Zhao, M. Zhang, K. Xu, Y. Zhang, X. Zhang, C. Hu, Association between PM1 exposure and lung function in children and adolescents: a systematic review and meta-analysis, IJERPH 19 (23) (2022) 15888, https://doi.org/10.3390/ijerph192315888.
[7] Y. Zhang, Z. Ding, Q. Xiang, W. Wang, L. Huang, F. Mao, Short-term effects of ambient PM1 and PM2.5 air pollution on hospital admission for respiratory diseases: case- crossover evidence from Shenzhen, China, Int. J. Hyg Environ. Health 224 (2020) 113418, https://doi.org/10.1016/j.ijheh.2019.11.001.
[8] L. Shi, X. Wu, M.D. Yazdi, D. Braun, Y.A. Awad, Y. Wei, P. Liu, Q. Di, Y. Wang, J. Schwartz, F. Dominici, M.A. Kioumourtzoglou, A. Zanobetti, Long-term effects of PM2.5 on neurological disorders in the American Medicare population: a longitudinal cohort study, Lancet Planet. Health 4 (12) (2020) e557–e565, https://doi.org/10.1016/S2542-5196(20)30227-8.
[9] J. Li, Y. Dong, Y. Song, B. Dong, A. van Donkelaar, R.V. Martin, L. Shi, Y. Ma, Z. Zou, J. Ma, Long-term effects of PM2.5 components on blood pressure and hypertension in Chinese children and adolescents, Environ. Int. 161 (2022) 107134, https://doi.org/10.1016/j.envint.2022.107134.
[10] S. Feron, R.R. Cordero, A. Damiani, R.B. Jackson, Climate change extremes and photovoltaic power output, Nat. Sustain. 4 (3) (2021) 270–276. https://www.nature.com/articles/s41893-020-00643-w.
[11] M.T. Udristioiu, L. Velea, R. Bojariu, S.C. Sararu, Assessment of urban heat Island for Craiova from satellite-based LST, AIP Conf. Proc. 1916 (1) (2017) 040004. doi: 10.1063/1.5017443.
[12] X. Hou, M. Wild, D. Folini, S. Kazadzis, J. Wohland, Climate change impacts on solar power generation and its spatial variability in Europe based on CMIP6, Earth Syst. Dynam. 12 (4) (2021) 1099–1113, https://doi.org/10.5194/esd-12-1099-2021.
[13] A. Mourad, A. Aissa, Z. Said, O. Younis, M. Iqbal, A. Alazzam, Recent advances on the applications of phase change materials for solar collectors, practical limitations, and challenges: a critical review, J. Energy Storage 49 (2022) 104186, https://doi.org/10.1016/j.est.2022.104186.

[14] Z. Song, J. Liu, H. Yang, Air pollution and soiling implications for solar photovoltaic power generation: a comprehensive review, App, Energy 298 (2021) 117247, https://doi.org/10.1016/j.apenergy.2021.117247.
[15] T. Alkharusi, G. Huang, C.N. Markides, Characterisation of soiling on glass surfaces and their impact on optical and solar photovoltaic performance, Renew. Energy 220 (2023) 119422, https://doi.org/10.1016/j.renene.2023.119422.
[16] T. Alkharusi, M.M. Alzahrani, C. Pandey, H. Yildizhan, Experimental investigation of nonuniform PV soiling, Solar Energy 272 (2024), https://doi.org/10.1016/j.solener.2024.112493.
[17] J. Gu, B. Yang, M. Brauer, K.M. Zhang, Enhancing the evaluation and interpretability of data-driven air quality models, Atmos. Environ. 246 (2021) 118125, https://doi.org/10.1016/j.atmosenv.2020.118125.
[18] T. Li, Q. Zhang, Y. Peng, X. Guan, L. Li, J. Mu, X. Wang, X. Yin, Q. Wang, Contributions of various driving factors to air pollution events: interpretability analysis from Machine learning perspective, Environ. Int. 173 (2023) 107861, https://doi.org/10.1016/j.envint.2023.107861. ISSN 0160-4120.
[19] X. Yan, Z. Zang, Y. Jiang, W. Shi, Y. Guo, D. Li, C. Zhao, L. Husi, A Spatial-Temporal Interpretable Deep Learning Model for improving interpretability and predictive accuracy of satellite-based PM2.5, Environ. Pollut. 273 (2021) 116459, https://doi.org/10.1016/j.envpol.2021.116459.
[20] K. Huang, Q. Xiao, X. Meng, G. Geng, Y. Wang, A. Lyapustin, D. Gu, Y. Liu, Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain, Environ. Pollut. 242 (A) (2018) 675–683, https://doi.org/10.1016/j.envpol.2018.07.016.
[21] Y. Ledmaoui, A.E. Maghraoui, M.E. Aroussi, R. Saadane, A. Chebak, A. Chehri, Forecasting solar energy production: a comparative study of machine learning algorithms, Energy Rep. 10 (2023) 1004–1012, https://doi.org/10.1016/j.egyr.2023.07.042.
[22] R. Opoku, G. Mensah, E.A. Adjei, J.B. Dramani, O. Kornyo, R. Nijjhar, M. Addai, Machine learning of redundant energy of a solar PV Mini-grid system for cooking applications, Sol. Energy 262 (2023) 111790, https://doi.org/10.1016/j.solener.2023.06.008.
[23] M.T. Udristioiu, Y.E. Mghouchi, H. Yildizhan, Prediction, modelling, and forecasting of PM and AQI using hybrid machine learning, J. Clean. Prod. 421 (2023) 138496, https://doi.org/10.1016/j.jclepro.2023.138496.
[24] P. Narkwatchara, C. Ratanatamskul, A. Chandrachai, Effects of particulate matters and climate condition on photovoltaic system efficiency in tropical climate region, Energy Rep. 6 (2020) 2577–2586, https://doi.org/10.1016/j.egyr.2020.09.016.
[25] Z. Song, M. Wang, H. Yang, Quantification of the impact of fine particulate matter on solar energy resources and energy performance of different photovoltaic Technologies, ACS Environ 2 (3) (2022) 275–286, https://doi.org/10.1021/acsenvironau.1c00048.
[26] T. Dewi, P. Risma, Y. Oktarina, A review of factors affecting the efficiency and output of a PV system applied in tropical climate, IOP Conf. Ser. Earth Environ. Sci. 258 (2019) 012039, https://doi.org/10.1088/1755-1315/258/1/012039.
[27] M. Fouad, L.A. Shihata, E.I. Morgan, An integrated review of factors influencing the performance of photovoltaic panels, Renew. Sustain. Energy Rev. 80 (2017) 1499–1511, https://doi.org/10.1016/j.rser.2017.05.141.
[28] S. Şevik, A. Aktaş, Performance enhancing and improvement studies in a 600 kW solar photovoltaic (PV) power plant; manual and natural cleaning, rainwater harvesting and the snow load removal on the PV arrays, Renew. Energy 181 (2022) 490–503, https://doi.org/10.1016/j.renene.2021.09.064.
[29] A.J. Barker, T.A. Douglas, E.M. Alberts, P.U.A. IreshFernando, G.W. George, J. B. Maakestad, L.C. Moores, S.P. Saari, Influence of chemical coatings on solar panel performance and snow accumulation, Cold Reg. Sci. Technol. 201 (2022) 103598, https://doi.org/10.1016/j.coldregions.2022.103598.
[30] A.D. Burakova, L.N. Burakova, I.A. Anisimov, O.D. Burakova, Evaluation of the operation efficiency of solar panels in winter, in: IOP Conference Series: Earth and Environmental Science, vol. 72, IOP Publishing, 2017 012022, https://doi.org/10.1088/1755-1315/72/1/012022, 1.
[31] L. Velea, M.T. Udriştioiu, S. Puiu, R. Motişan, D. Amarie, A community-based sensor network for monitoring the air quality in urban Romania, Atmosphere 14 (5) (2023) 840, https://doi.org/10.3390/atmos14050840.
[32] M.T. Udristioiu, L. Velea, R. Motisan, First results given by the independent air pollution monitoring network from Craiova city Romania, AIP Conf. Proc. 2843 (2023) 040001, https://doi.org/10.1063/12.0017678.
[33] S.S. Skiena, The Data Science Design Manual, Springer, 2017, 978-3-319- 55443-3.
[34] J. Rabčan, P. Rusńak, S. Subbotin, Classification by Fuzzy decision trees Inducted based on Cumulative Mutual information, in: Proceedings of 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), 2018, pp. 208–212, 978-1-5386-2556-9.
[35] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2019. Published independently, ISBN 979-8411463330.
[36] B. Rozemberczki, L. Watson, P. Bayer, H.T. Yang, O. Kiss, S. Nilsson, R. Sarkar, The Shapley Value in Machine Learning, 2022 arXiv preprint arXiv:2202.05594.
[37] R. Rodríguez-Pérez, J. Bajorath, Interpretation of machine learning models using Shapley values: application to compound potency and multi-target activity predictions, J. Comput. Aided Mol. Des. 34 (2020) 1013–1026, https://doi.org/10.1007/s10822-020-00314-0.
[38] M. Zareba, H. Dlugosz, T. Danek, E. Weglinska, Big-data-driven machine learning for enhancing spatiotemporal air pollution pattern analysis, Atmosphere 14 (4) (2023) 760, https://doi.org/10.3390/atmos14040760.