energies

MDPI

Article

Machine Learning Methods to Forecast the Concentration of PM10 in Lublin, Poland

Justyna Kujawska [1], Monika Kulisz [2,*], Piotr Oleszczuk [2] and Wojciech Cel [1]

[1] Faculty of Environmental Engineering, Lublin University of Technology, 20-618 Lublin, Poland
[2] Faculty of Management, Lublin University of Technology, 20-618 Lublin, Poland
* Correspondence: m.kulisz@pollub.pl; Tel.: +48-669-428-542

Abstract: Air pollution has a major impact on human health, especially in cities, and elevated concentrations of PMx are responsible for a large number of premature deaths each year. Therefore, the amount of PM10 in the air is monitored and forecasts are made to predict the air quality. In Poland, mainly deterministic models are used to predict air pollution. Accordingly, research efforts are being made to develop other models to forecast the ambient PM10 levels. The aim of the study was to compare the machine learning models for predicting PM10 levels in the air in the city of Lublin. The following machine learning models were used: Linear regression (LR), K-Nearest Neighbors Regression (KNNR), Support Vector Machine (SVM), Regression Trees (RT), Gaussian Process Regression Models (GPR), Artificial Neural Network (ANN) and Long Short-Term Memory network (LSTM). The collected data for three consecutive years (January 2017 to December 2019) were used to develop the models. In total, 19 parameters, covering meteorological variables and concentrations of several chemical species, were explored as potential predictors of PM10. The data used to build the models did not take into account the seasons. The algorithms achieved the following R^2 values: 0.8 for LR, 0.79 for KNNR, 0.82 for SVM, 0.77 for RT, 0.89, 0.90 for ANN and 0.81 for LSTM. Research has shown that the selection of a machine learning model has a large impact on the quality of the results. In this research, the ANN model performed slightly better than other models. Then, an ANN was used to train a network with five output neurons to predict the approximate level of PM10 at different time points (PM level at a given time, after 1 h, after 6 h, after 12 h and after 24 h). The results showed that the developed and tuned ANN model is appropriate (R = 0.89). The model created in this way can be used to determine the risk of exceeding the PM10 alert level and to inform about the air quality in the region.

Keywords: air pollution; ANN; PM10 forecasting; air quality modeling

Citation: Kujawska, J.; Kulisz, M.; Oleszczuk, P.; Cel, W. Machine Learning Methods to Forecast the Concentration of PM10 in Lublin, Poland. *Energies* **2022**, *15*, 6428. https://doi.org/10.3390/en15176428

Academic Editors: Esmaeel Eftekharian and Robert H. Ong

Received: 27 July 2022
Accepted: 31 August 2022
Published: 2 September 2022

1. Introduction

Air pollution is a global problem. In Europe, ambient air quality remains poor in many areas, particularly in urban ones [1]. Growing industrialization contributes to an increase in air pollutants, such as sulfur dioxide (SO_2), particulate matter (PM10), carbon dioxide (CO_2), ozone (O_3), nitrogen oxide (IV) (NO_2), nitrogen oxygen (NO_x), carbon monoxide (II) (CO), benzene (C_6H_6), etc. Due to the increase in industrialization, air pollutants are on the rise, negatively affecting human health and nature, causing danger to human life. Automobile transportation and fuel combustion in the residential and commercial sectors are also responsible for the increase in air pollution. In addition to air pollution from anthropogenic sources, there are those from natural sources, such as grass pollination, soil erosion and rock weathering [2–4]. Therefore, air pollution is closely monitored and analyzed at measuring stations. The stations are located in regional zones, designated according to the aerodynamic condition. The location criteria and requirements for the quality of measurements and other methods of assessing air quality are specified in the Regulation of the Minister of Climate and Environment of 11 December 2020 on assessing

levels of substances in the air (Journal of Laws 2020, item 2279 [5]. Reference measurements can be divided into gravimetric manual measurements (PN-EN 12,341 standard) and automatic measurements (PN-EN 16,450 standard) [6–8].

Out of these pollutants, PM particles are the most abundant and varied, playing a major role in assessing the impact on health and the environment. According to the Air Quality Framework Directive, the PM10 definition is as follows: "PM10 shall mean particulate matter which passes through a size-selective inlet with a 50% efficiency cut-off at 10 µm aerodynamic diameter" [9]. PM10 particles include sulfur, heavy metals, highly toxic chemical organic compounds, such as dioxins and polycyclic aromatic hydrocarbons (e.g., benzo-a-pyrene), and allergens, including pollen and fungal spores, among others. The main sources of particulate matter are the cement, fertilizer, ceramic, chemical, wood and energy industries, as well as transportation and municipal sources: from households, landfilling and waste disposal [8,10–13]. In addition, the chemical composition of PM10 particles depends on the source of origin, which makes it difficult to protect human health. For this reason, a maximum acceptable level was defined in the national air quality objective system. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on air quality and cleaner air for Europe imposes two standards when it comes to the maximum permissible concentrations of PM10 [14]. The first relates to annual average concentrations. The maximum permissible annual average concentration of PM10 in the air is 40 $\mu g/m^3$. A standard for the 24-h average concentration was also established at 50 $\mu g/m^3$; additionally, it was noted that the standard for the daily average concentration can be exceeded for a maximum of 35 days per year. However, it should be remembered that the World Health Organization guidelines are stricter. According to the WHO, the average annual concentration of PM10 should not exceed 20 $\mu g/m^3$ [15]. Therefore, all countries are required not to exceed the permissible concentrations of PM10.

Poland is a country with poor air quality. WHO reports that among the 50 cities of the European Union, as many as 33 are located in Poland. Despite the observed reduction in emissions of particulate matter precursors (especially sulfur dioxide) and measures taken to reduce concentrations of particulate matter in the air, high concentrations of PM10 and PM2.5 particulate matter remain the most significant air quality problem in Poland. In Poland, exceedances of the permissible values of PM10 particulate matter concentrations generally occur in winter. These exceedances are primarily related to particulate matter emissions from individual heating of buildings (45%) and transportation (5%). Next are emissions from industrial plants, heating plants, power plants (12%) and unfavorable meteorological conditions [16]. In the 2019 air quality assessment for PM10, of the 46 zones subject to assessment based on 24-h concentrations, only 12 zones recorded concentration levels that did not exceed permissible values. The remaining zones (34) show that the concentration levels are above the permissible level [17].

Cities in Poland, including the largest agglomerations, show very high diversity in terms of air quality [11]. Połednik (2022) analyzed the air pollutant emissions in the last four years (2018–2021) in the Upper Silesian Region, which has one of the worst air qualities in Poland and in Europe in general, and the emissions in the Lublin Region in eastern Poland, which is considered a clean region. The obtained results indicated that in both agglomerations, the exposure to air pollutants was on similar levels, which were several times higher than in the remaining parts of the considered regions and the average values for Poland in general [1]. In 2018, in the Lublin Agglomeration zone, the PM10 content ranged from 27.6 to 33.7 $\mu g/m^3$ at all measurement sites. The largest fraction of PM10, in total suspended particulate (TSP) in the Lublin Voivodeship, is found in the municipal and residential sectors. A significant portion of the PM10 emissions from road transport comes from processes other than fuel combustion, which include abrasion of vehicle tires and brakes as well as abrasion of road surfaces. According to the 2018 Annual Assessment of Air Quality in the Lublin Voivodeship, the Lublin Agglomeration zone was classified as Class C, in terms of PM10 suspended particulate matter concentrations. This forced

the provincial authorities to develop an Air Protection Program as well as identify the corrective measures that will lead to an improvement in air quality status [18].

For this reason, the PM10 levels are continuously monitored and models for forecasting as well as estimating the PM10 concentrations are created based on continuous and periodic PM10 measurement results from monitoring stations. The obligation to perform short-term forecasts of air pollution is related to the need to determine the risk of exceeding the alert, permissible or target level of substances in the air for the purpose of informing the public as well as the local provincial crisis management centers and provincial boards in accordance with Article 94 (paragraph 1b) and (paragraph 1c) of the Law on Environmental Protection [19]. Simultaneously, this obligation constitutes the implementation of one of the main objectives of Directive 2008/50/EC of the European Parliament and of the Council [14]. The purpose of the task is to provide the information on the forecast concentrations of air pollutants. This information is necessary to warn the public about the risk of high, health-threatening concentrations of pollutants, as well as to trigger the actions provided for short-term action plans, in accordance with the requirements of the aforementioned directive. The European Union directives on air quality do not specify specific models, simply defining targets and accuracy requirements. It should be mentioned that in most European countries, a wide range of air quality models are employed, not only deterministic ones [20]. In Poland, the following deterministic models have been mainly used for air quality forecast modeling for several decades: non-diffusive box model, diffusive Gauss (plume or cloud) model and diffusive CTM (Chemical Transport Model) models. In all these models, the input is an emission map and meteorological data. Modeling is performed in a computational grid, so it is necessary to generalize the various physical features of the area to the mesh size of the models. When the resolution of the calculation is too low, this generalization can be a source of error [21].

Extensive and complicated deterministic models for calculating air quality forecasts have prompted the search for faster forecasting models and the use of numerical models for air quality forecasting [22].

On the basis of a literature analysis, it can be concluded that the numerical models being developed for forecasting local levels of particulate matter in the air have the ability to relate complex relationships between input and output variables based on the measurement data from monitoring stations and have good forecasting quality [23,24]. For example, Arhami et al. (2013) [25] developed an ANN model for forecasting hourly criteria of pollutants (NOx, NO_2, NO, O_3 and PM10) in Tehran using only meteorological data as input variables, i.e., wind direction, wind speed, relative humidity and air temperature. The results showed that appropriate ANN models can be used as reliable metamodels for the prediction of hourly air pollutants in urban environments. High correlations were obtained with R^2 of more than 0.82 between the modeled and observed hourly pollutant levels for CO, NOx, NO_2, NO and PM10. However, the predicted O_3 levels were less accurate [25]. Suleiman et al. (2019) presented a method for evaluating the effectiveness of roadside PM10 and PM 2.5 reduction scenarios using: Artificial Neural Network (ANN), Boosted Regression Trees (BRT) and Support Vector Machines (SVM). All models performed very well in predicting the concentrations of PM10 with around 95% of their predictions, falling within the factor of two of the observed concentrations at the roadsides. The results show that the BRT and SVM models for PM10 predictions performed slightly better than the ANN models, as indicated by the smaller *RMSE* values (7.99 and 7.72) [26]. Mehdipour et al. (2018) applied three different artificial intelligence models: Bayesian network (BN), support vector machines (SVM) and decision tree (DT) to predict PM in Tehran. The model input parameters were temperature, rainfall, wind speed, nebulosity, relative humidity, insolation, O_3, PM10, SO_2, NO_2 and CO. The SVM model showed the highest correlation coefficient for the modeled data and observed data was 0.9414, compared to the other models tested, i.e., DT and BN, for which the correlation coefficients were 0.92046 and 0.8927, respectively [27].

Krishan et al. (2019) used meteorological data, transportation emissions and traffic data as input to model hourly concentrations of air quality indicators: concentrations of

O_3, PM10, NOx and CO in Delhi, India using Long Short-Term Memory network (LSTM). Performance evaluation of LSTM algorithms for hourly concentration prediction was carried out during 2008–2010, and it was found that LSTM models efficiently deal with the complexities and is immensely effective in ambient air quality forecasting LSTM models performed quite well for all the four variables, achieving high correlation coefficients (0.92–0.98) [28].

Artificial intelligence models show good forecasting quality for air quality indicators [28,29]. This is because these models are capable of handling multidimensional input data, and do not require data preprocessing for input parameters. The advantage of the afore-mentioned models (ANN, SVM, etc.) is that they have a higher ability to predict the air quality parameters than in deterministic, empirical methods and linear regression models. In turn, the disadvantage is that these models give a prediction of concentrations only for measurement points and assume invariability of emissions over a certain time scale. Therefore, combining and analyzing different models can provide the results with lower error variance compared to single models [29,30].

This research focuses on verifying the applicability of numerical models for PM10 forecasting from measurement database data. The measurement data from the Radawiec station, located in Lublin, a city in eastern Poland, was used for the study. There is a lack of articles in the literature on air quality prognostication in eastern Poland, so the city of Lublin was chosen for air quality forecasting [31,32]. The lack of air quality models in Lublin is most likely due to the fact that Lublin is a city with good and even very good air quality.

The new approach proposed in this article is to create models using measured air pollution–meteorological and chemical data—without considering the seasons. In Poland, most of the machine learning models created so far are based only on meteorological data and consider only the winter season [31]. In the present paper, the data from air quality monitoring stations were used to predict the PM10 levels.

The following meteorological data from air quality monitoring stations were used as inputs to build models: temperature (T), relative humidity (RH), wind speed (WS), wind direction (WD) and chemical air pollution: SO_2, PM10, NO_2, NO_x, CO, O_3 and C_6H_6 as well as the effect of measurement hour (h) and month (M).

The first stage of the research was to analyze selected machine learning methods (Linear regression (LR), K-Nearest Neighbors Regression (KNNR), Support Vector Machine (SVM), Regression Trees (RT), Gaussian Process Regression Models (GPR), Artificial Neural Network (ANN) and Long Short-Term Memory network (LSTM)). The results of these models were analyzed to find the best fit for predicting the PM10 levels. The quality of these models was compared and evaluated, taking into account the Mean Squared Error (*MSE*), determination coefficient (R^2) and the regression coefficient (*R*). The best model was then selected and this method was used to model the approximate level of PM10 at different time points (at a given time, after 1 h, after 6 h, after 12 h and after 24 h). A novelty is the modeling of a network with five outputs to simultaneously predict the approximate level of PM10 at these time points.

Forecasting at different time points is important in terms of determining the risk of exceeding PM10 alert and air quality information levels in a given region. Such models can provide an alternative to those currently in use.

This paper consists of five sections. The first section, the Introduction, contains a description of the problems for forecasting air pollution. The second section, entitled Materials and Methods, describes the procedure for acquiring measurement data by means of the machine learning models used in the research. The third section includes the research results obtained by using LR, KNNR, SVM, RT, GPR, ANN and LSTM. The fourth section described the ANN model for predicting PM10 at different time points. The fifth section is a discussion, taking into account the most important aspects of the analyses carried out in this area and the results obtained. The final section provides a summary and conclusions as well as the information on future research.

2. Materials and Methods

2.1. The Dataset

Results from the measurement station included in the National Environmental Monitoring network, recorded in the period between 2017 and 2019, were taken into account and the Air Quality Portal of the Chief Inspectorate for Environmental Protection was the source of data used for modeling. The dataset contains both meteorological and air pollution data read hourly in the Radawiec commune in the Lublin Voivodeship, Poland (51.21304 N, 22.385393 E), in 2017–2019. The analysis took into account data from an urban background station (located within agglomeration boundaries). Reference measurements were made automatically in accordance with the PN-EN 16,450 standard. The meteorological data included: temperature (T), relative humidity (RH), wind speed (WS) and wind direction (WD), whereas the air pollution data involved: SO_2, PM10, NO_2, NO_x, CO, O_3 and C_6H_6. The entire dataset, after removing missing data, contained a total of 23,300 hourly readings. In addition, the effect of measurement hour (h) and month (M) was also included in the conducted research, as they can have a large impact on the level of month and hour, and together with the wind direction, they are circular in nature. For this reason, a new two-dimensional variable was introduced for each of them, being its sine and cosine: sin_h, cos_h, sin_M cos_M, sin_WD, cos_WD. The descriptive statistics of the measured data are presented in Table 1.

Table 1. Descriptive statistics of the data.

	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
SO_2	0	56.9	4.93	3.47	3.13	20.13
PM10	0.5	496	30.95	25.94	4.79	46.67
NO_2	0	128.2	21.07	15.21	1.77	4.23
NO_x	0	766.7	32.21	39.79	6.52	70.76
CO	0	5.32	0.36	0.3	4.76	41.40
O_3	0	169.6	48.88	28.31	0.41	−0.30
C_6H_6	0.05	25.3	1.69	1.5	4.52	42.40
T	−15.34	36.84	10.22	9.37	0.10	−0.85
RH	21	100	72.39	18.51	−0.44	−0.95
WS	0	20.67	5.21	2.87	1.19	1.75
WD	0.65	360	196.66	94.99	−0.28	−0.99

2.2. Methods

To predict the approximate level of PM10, several machine learning methods have been tested: Linear regression (LR), K-Nearest Neighbors Regression (KNNR), Support Vector Machine (SVM), Regression Trees (RT) and Gaussian Process Regression Models (GPR). Linear regression was also used as a tool to investigate the selection of input parameters for the Artificial Neural Network (ANN) and Long Short-Term Memory network (LSTM) model.

Many algorithmic methods have been preliminarily analyzed. Ultimately, the following models based on machine learning were selected: LR, KNNR, SVM, RT and GPR, as well as neural networks—shallow, with one hidden layer (ANN), and deep, consisting of multiple hidden layers (LSTM). Among regression models, LRs are popular and are used first, but the highly restricted form of these models means that they often have low predictive accuracy. Therefore, more flexible models, such as RTs or SVMs, KNNs or GPRs are developed. RTs are easy to interpret and allow for fast fitting and prediction and low memory consumption; moreover, linear SVMs are easy to interpret but can have low predictive accuracy, while non-linear SVMs are more difficult to interpret but can be more accurate, such as KNN. Among these methods, GPR is characterized by high accuracy. All of the aforementioned methods use supervised learning, since the input and output parameters of the phenomenon being modeled are known.

The best-performing model was used to predict the PM10 level at different time points (at a given time, after 1 h, after 6 h, after 12 h and after 24 h). The creation of an ANN model with five output neurons for predicting ambient PM10 concentrations is essential

for effective air quality management and the development of air quality-related policies. The model created in this way can find application in public warning systems, indicating the situations that could potentially cause direct threats to human health in as many as five time periods. The modeling process was performed in Matlab R2022a (The MathWorks, Inc., Natick, MA, USA) [33] and R 4.1.2 (R Foundation for Statistical Computing, Vienna, Austria) [34] environments. Modeling was performed on a computer with the following parameters: AMD Ryzen™ 7 5800H (8 cores, 16 threads, 3.20–4.40 GHz, 20 MB cache), 16 GB RAM, NVIDIA GeForce RTX 3060 graphics card and AMD Radeon™ Graphics, graphics card memory—6 GB GDDR6. The study was conducted in major steps, as shown in Figure 1.

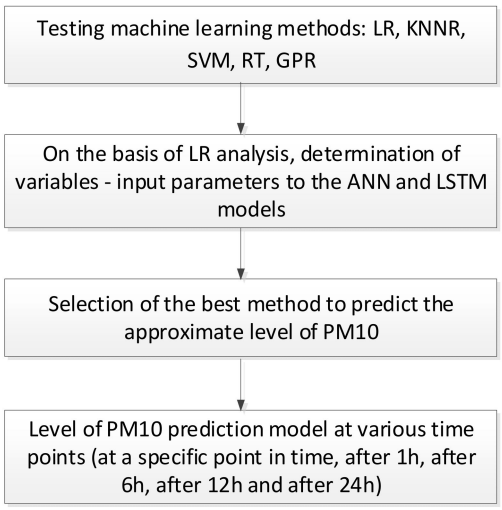


Figure 1. Scheme of the methodology used to conduct the research.

The training set was 70% of the dataset, and the test set constituted the remaining 30%. Measures of goodness of fit of the model used in this study were Mean Squared Error (MSE) and determination coefficient (R^2). The Mean Squared Error is defined by:

$$MSE = \frac{1}{n}\sum_{n=1}^{n}(\hat{y}_i - y_i)^2$$

where y_i is the actual value of the PM10 level and $\hat{y}_i$ denotes the value of the PM10 level for the i-th observation obtained from the model. The determination coefficient is given by the formula:

$$R^2 = \frac{\sum_{n=1}^{n}(\hat{y}_i - \overline{y_i})^2}{\sum_{n=1}^{n}(y_i - \overline{y_i})^2},$$

while the regression coefficient R measures the correlation between outputs and inputs and was calculated according to the formula:

$$R(y', y^*) = \frac{cov(y', y^*)}{\sigma_{y'}\sigma_{y^*}} \qquad R \in <0,1>$$

where $\sigma_{y'}$ is the standard deviation of the reference values and σ_{y^*} is the standard deviation of the predicted values. The mean absolute error (MAE) indicator was also used to compare the models, which is calculated as:

$$MAE = \frac{1}{n}\sum_{n=1}^{n}|\hat{y}_i - y_i|,$$

where $|\hat{y}_i - y_i|$ are the absolute errors. The root mean square error (*RMSE*)I is calculated according to the formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{n=1}^{n}(\hat{y}_i - y_i)^2}.$$

The higher the regression coefficient R and R^2 and the lower *MSE*, *RMSE* and *MAE*, the better the quality of the generated models [35].

The first step was linear regression. The simplest model has the following form:

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k + \varepsilon, \quad (1)$$

where $X_1, X_2, \ldots, X_n$ are independent variables (predictors), Y denotes dependent variable, i.e., the PM10 level in the considered case, $b_1, b_2, \ldots, b_k$ are model coefficients and ε is random component (model residuals). It allows analyzing the influence of independent variables (meteorological, air pollution variables, etc.), expressed in model parameters, on the dependent variable i.e., PM10 level. In addition to the basic model, the following variants of linear regression were examined: model with interactions (basic linear model with additional interaction terms being products of predictors), pure quadratic (basic linear model with purely quadratic terms of the predictors) and quadratic (basic linear model with both interaction and pure quadratic terms). The following model fitting methods were used: least squares, robust (modified objective function to make model less sensitive to outliers) and stepwise regression. In the case of Least Squares linear Regression, the quality of the model was verified by checking the fulfillment of the following assumptions: the normality of the residual distribution (Shapiro–Wilk test), the residual homoscedasticity (Breusch–Pagan test), the lack of autocorrelation of the residuals (Durbin–Watson test) and the existence of outliers (Cook distance, diagnostic plots).

In the next step the mentioned machine learning methods were tested: KNNR, SVM, Regression Trees, GPR, ANN and LSTM. The parameters used during modeling for these methods are shown in Table 2.

Table 2. Modeling parameters using the analyzed methods.

Methods	Model Parameters
K-Nearest Neighbor Regression (KNNR)	The dataset was normalized and the Euclidean distance was used to find the closest neighbors, k = 1, 2, . . . , 10 were tested.
Support Vector Machine (SVM)	Various Kernel functions were employed for training SVM: Gaussian kernel, Linear kernel, Quadratic kernel, Cubic kernel. Kernel scale, box constraint, epsilon—Automatic, standardize data: true.
Regression Trees (RT)	Minimum leaf size setting was changed while training RT. The analysis was conducted using Minimum leaf size—4, 12 and 36. Surrogate decision splits—Off.
Gaussian Process Regression Models (GPR)	GPR was trained using various Kernel functions: Rational Quadratic, Squared Exponential, Matern 5/2 and Exponential. Hyperparameters: basis function: Zero, Constant and Linear, use isotropic kernel: true, kernel scale, signal standard deviation and sigma: Automatic, standardize, optimize numeric parameters: true
Artificial Neural Network (ANN)	Three different algorithms were used to train the network: the Levenberg–Marquardt algorithm, Bayesian regularization algorithm and Scaled conjugate gradient algorithm. The number of neurons in the hidden layer (10–300) was selected experimentally. In this case, the learning set was 70%, whereas the test and validation sets were 15% each. Networks were built with one hidden layer.
Long Short-Term Memory network (LSTM)	To teach the network, the number of hidden units was experimentally selected in the range of 500 to 2000. Solver for training network—'Adam', dropout layers—0.2, mini-batch size—changed in the range of 500–1000, option to pad, truncate, or split input sequences, specified as longest. The learning set for this network was 70%, and the test and validation sets were 15% each.

3. Results

3.1. Models Obtained Using Machine Learning Methods

Table 3 presents model quality parameters obtained using the following models: LR, KNNR, SVM, RT and GPR. Linear regression was also a tool to investigate the selection of input parameters for ANN and LSTM.

Table 3. Quality parameters of the models obtained using machine learning methods.

Model Quality Parameters	Models Obtained Using Machine Learning Methods				
	LR	**KNNR**	**SVM**	**RT**	**GPR**
R^2	0.8	0.79	0.82	0.77	0.89
MSE	135.51	135.24	119.3	156.57	85.36

For the Linear Model, the coefficient of determination (R^2) of the basic linear regression model was equal 0.73, while the mean square error (*MSE*) was 165.57. Verification of assumptions showed that the model met no normal distribution of residuals and the lack of residual homoscedasticity. The stepwise regression method did not result in producing a better model that met the assumptions. The best results were obtained for Linear Model with interactions and R^2 and *MSE* were equal to 0.8 and 135.51.

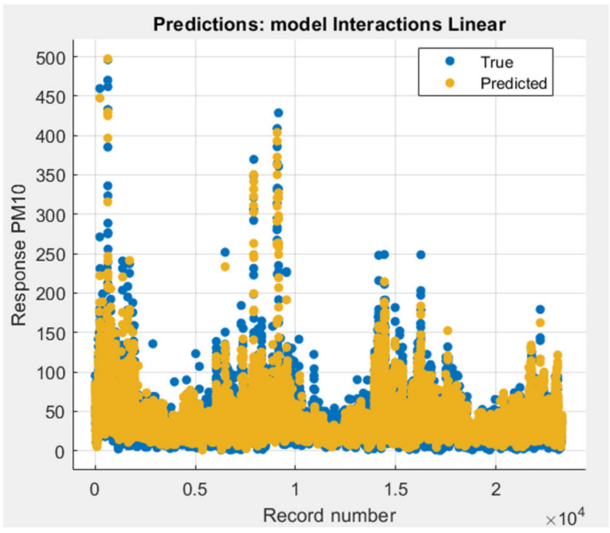
For K-Nearest Neighbor Regression (KNNR), the best result was obtained for $k = 2$. In this case, the R^2 and *MSE* values were equal to 0.79 and 135.24, respectively. For the Support Vector Machine (SVM) the best results were obtained with Cubic kernel ($R^2 = 0.82$ and $MSE = 119.3$); for Regression Trees with a minimum leaf size—12 ($R^2 = 0.77$ and $MSE = 156.57$)—and for Gaussian Process Regression Models (GPR) with Exponential kernel and basis function: constant ($R^2 = 0.9$ and $MSE = 65.36$).

In addition, Figure 2 shows the response plot and the predicted response versus the true response plots for the models presented in Table 3. The response plot plots are for the LR (Figure 2a), KNNR (Figure 2c), SVM (Figure 2e), RT (Figure 2g) and GPR (Figure 2i) models, respectively, while the predicted response versus true response plots are for the LR (Figure 2b), KNNR (Figure 2d), SVM (Figure 2f), RT (Figure 2h) and GPR (Figure 2j) models, respectively. The response plot shows both the true and predicted responses. The predicted response versus true response charts can be used to understand how well the regression model makes predictions for different response values.

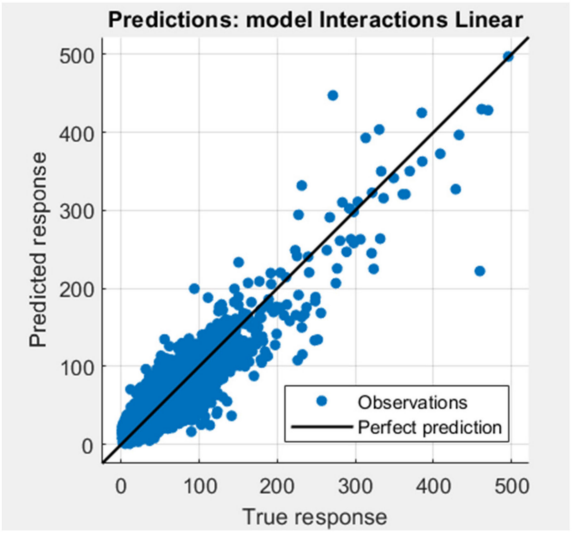
3.2. Artificial Neural Network Model

The next stage of the conducted research was to predict the approximate level of PM10 using ANN models. In this study, the selection of variables was based on linear regression analysis. All parameters, i.e., temperature (T), relative humidity (RH), wind speed (WS), wind direction (WD) and air pollution data, including SO_2, PM10, NO_2, NO_x, CO, O_3, C_6H_6, sin_h, cos_h, sin_M cos_M, sin_WD and cos_WD, were used for neural network modeling.

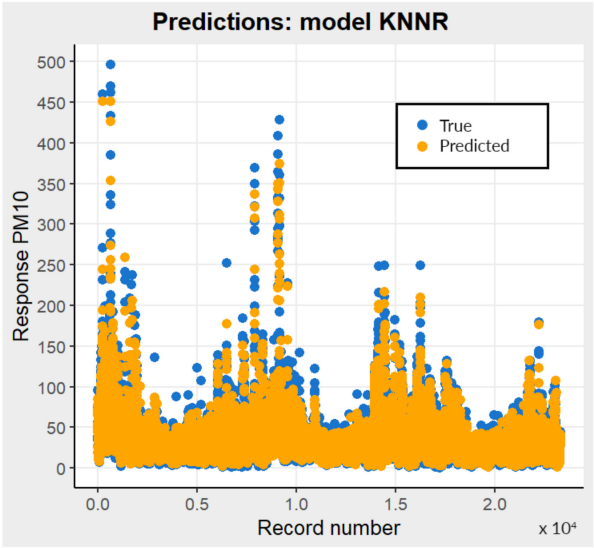
The best modeling results were obtained for a network with 220 neurons, using the Levenberg–Marquardt algorithm, which was obtained in 27 iterations. A schematic of the network is shown in Figure 3. Other data, including performance validation and the rate of error decrease (gradient) and Mu, are presented in Table 4. In general, during modeling, the error decreases after successive learning periods, but may begin to increase in the validation dataset when the network begins to over-fit the learning data. Learning stops after six consecutive increases in validation error (or no decrease in error), and the best results are obtained from the iteration with the lowest validation error. The best validation performance was obtained for iteration 21, as shown in Figure 4, while Figure 5 shows the early stopping strategy.



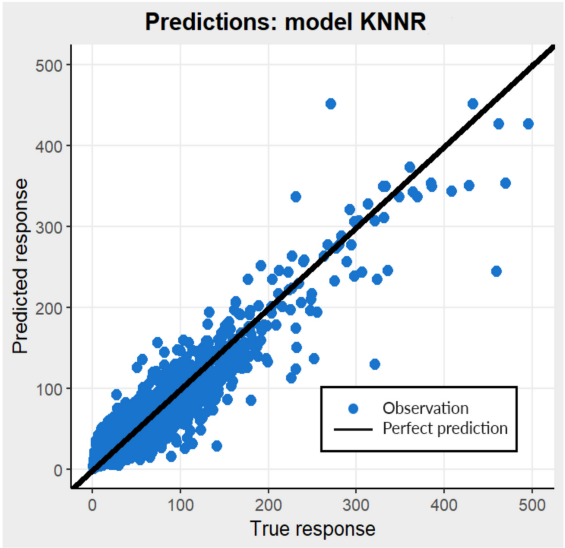
(**a**)



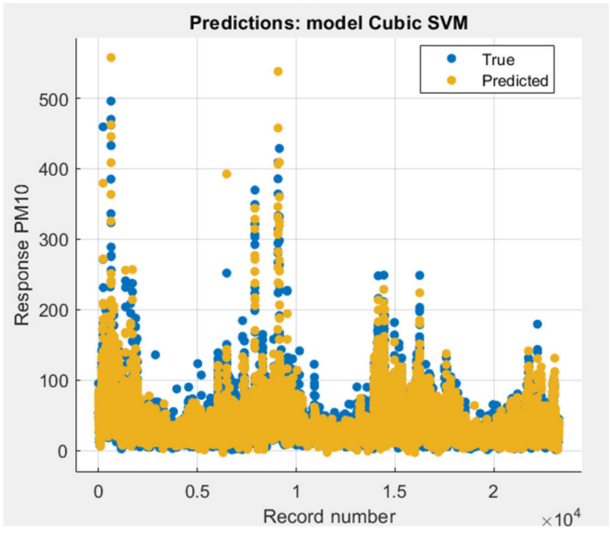
(**b**)



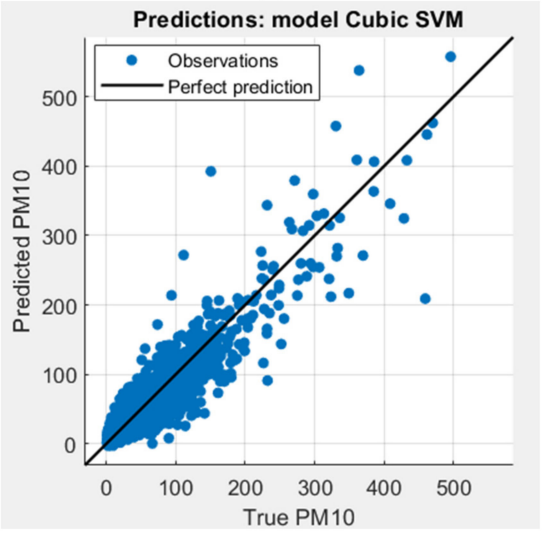
(**c**)



(**d**)



(**e**)



(**f**)

Figure 2. *Cont.*

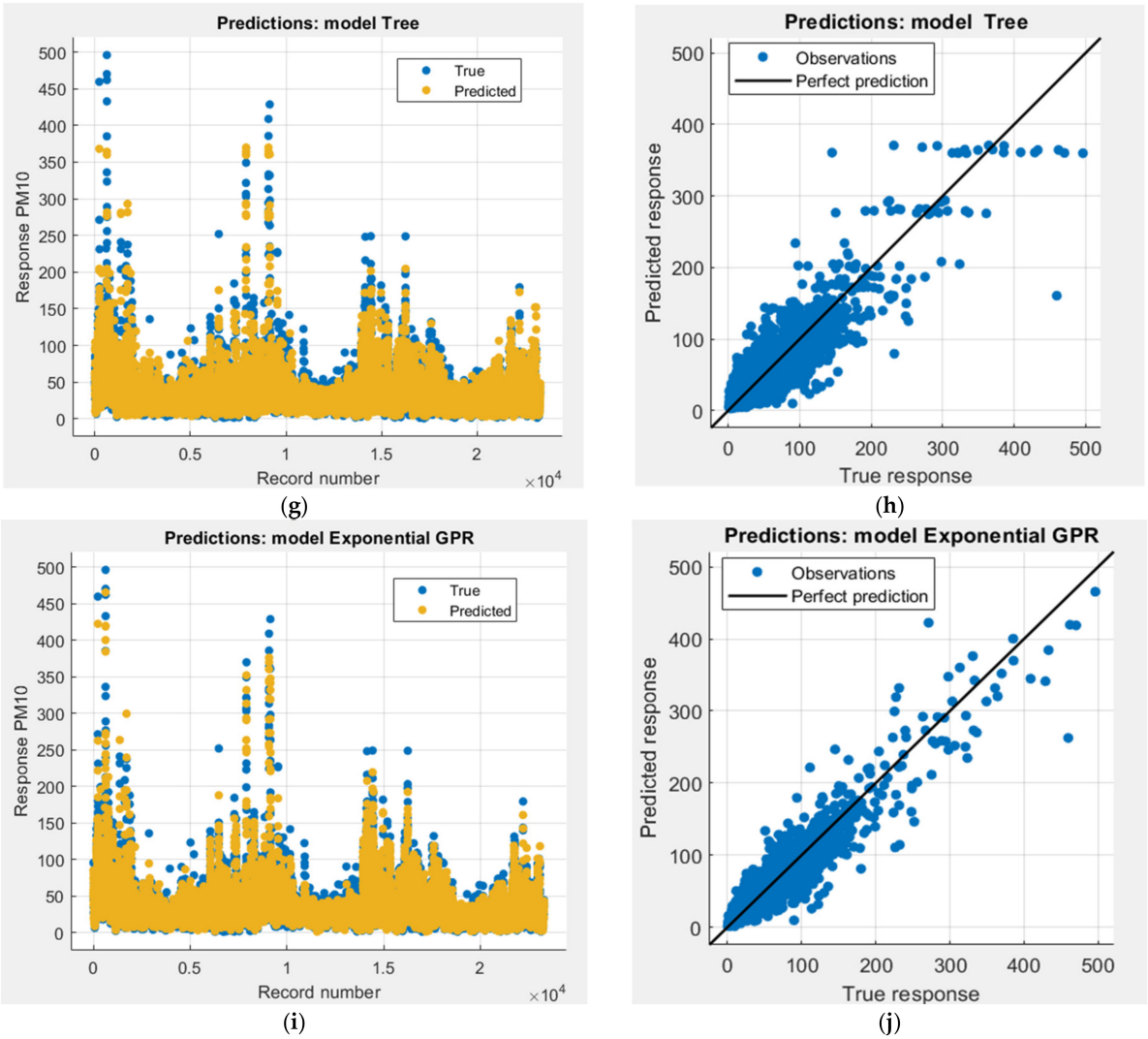


Figure 2. Response plot graphs for the models (**a**) LR, (**c**) KNNR, (**e**) SVM, (**g**) RT and (**i**) GPR, and the predicted response versus true response graphs for the models (**b**) LR, (**d**) KNNR, (**f**) SVM, (**h**) RT and (**j**) GPR, respectively.

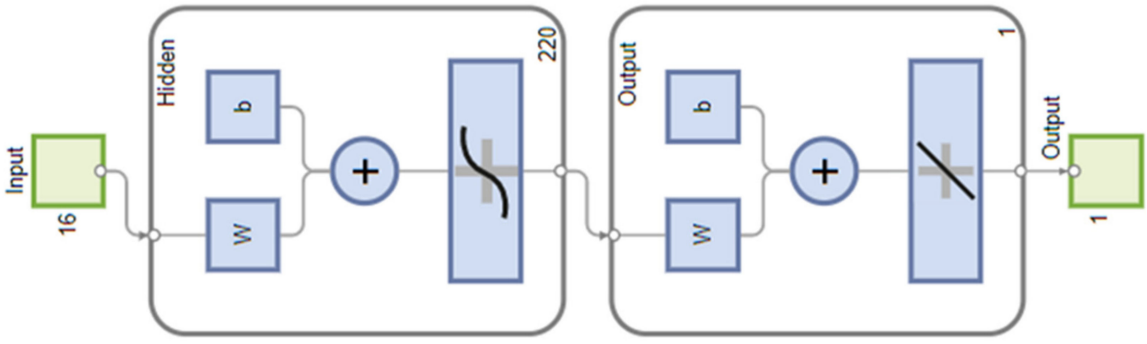


Figure 3. The ANN structure.

Table 4. Network training conditions.

Unit	Initial Value	Stopped Value	Target Value
Epoch	0	27	1000
Elapsed Time	-	00:06:49	-
Performance	2.63×10^6	44.9	0
Gradient	8.39×10^6	84	1×10^{-7}
Mu	0.001	0.01	1×10^{10}
Validation Checks	0	6	6

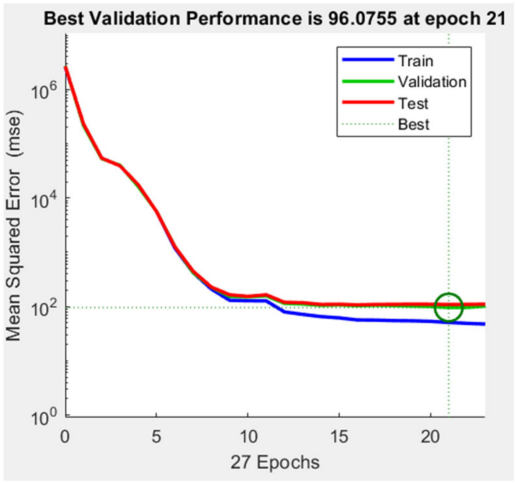


Figure 4. ANN teaching performance.

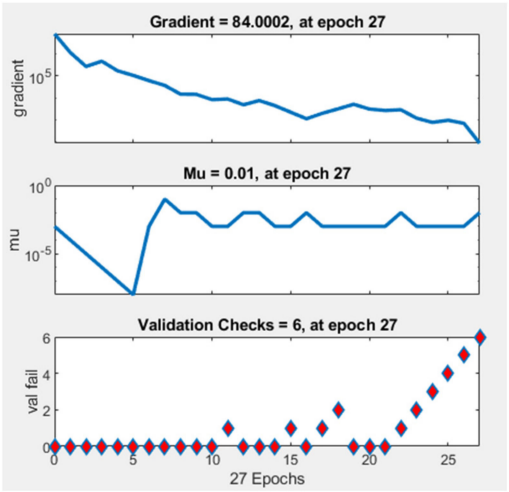


Figure 5. Early stopping strategy for the ANN model.

Table 5 shows the network learning results (*MSE*, R^2 and Regression *R* value) by learning, validation and test subsets. In addition, regression statistics are shown in Figure 6, for which the regression (*R*) value for the learning data is 0.96, and 0.94 for the validation data and 0.92 for the test data. The overall regression was 0.95, which represents the degree of overlap between the measurement points and the fit line with the ideal prediction line $Y = T$.

Table 5. ANN regression statistics and *MSE*.

	Observations	*MSE*	*R*	R^2
Training	16,310	50.93	0.96	0.92
Validation	3495	96.07	0.94	0.86
Test	3495	109.11	0.92	0.83

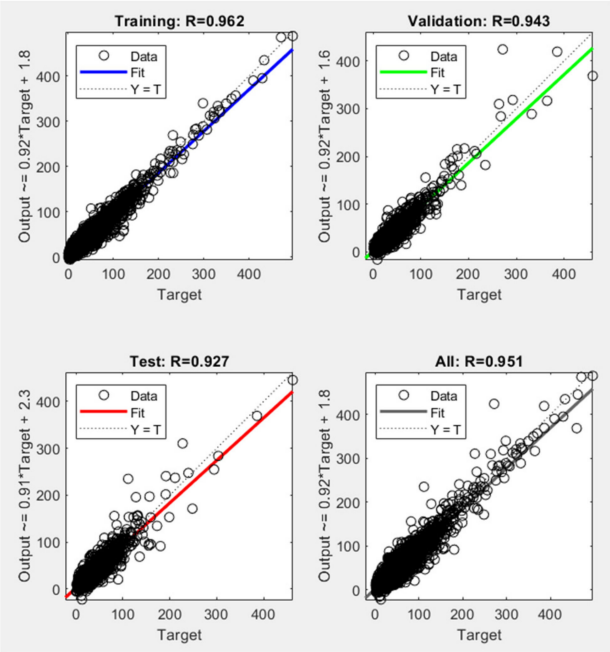


Figure 6. ANN regression statistics for individual sets and the total set.

3.3. Long Short-Term Memory Network Model

The last model analyzed was the LSTM model, for which, as with the ANN models, all input parameters were used. The best modeling results were obtained for the network with the number of 2000 epochs, with Mini-Batch Size = 1000. The network diagram is shown in Figure 7. The rest of the network learning data is shown in Figure 8. The number of iterations per epoch was 1. Learning was done at a constant rate of 0.001. This property was established to make the learning process more accurate. The regression layer is the last in the considered model. For typical regression problems, the regression layer must follow after the final fully connected layer (fc).

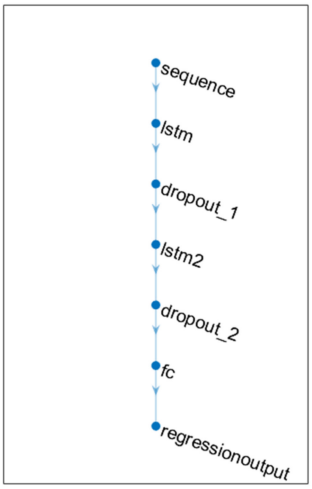


Figure 7. The LSTM structure.

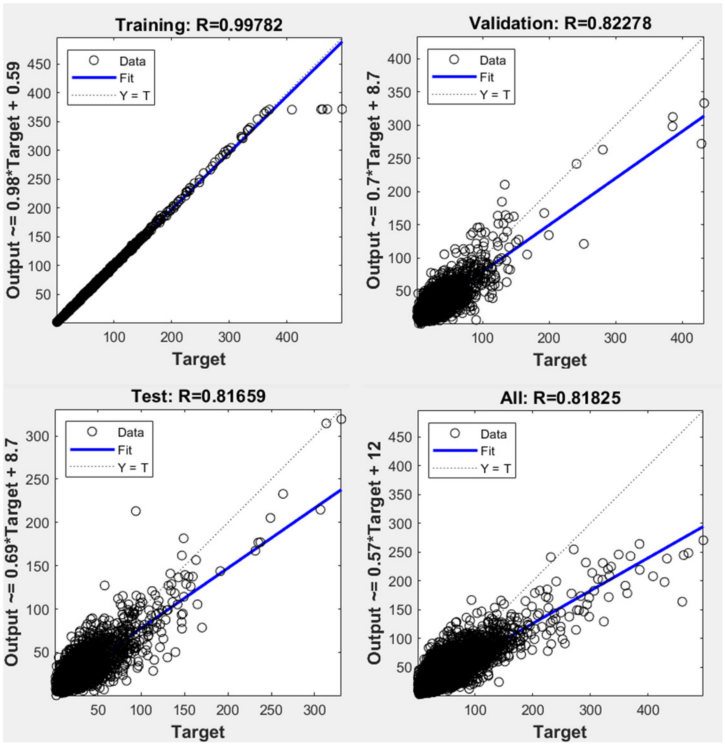


Figure 8. LSTM regression statistics for individual and the total set.

Table 6 shows the network learning results (*MSE*, R^2 and Regression *R* value) by learning, validation and test subsets. In addition, regression statistics are shown in Figure 9, for which the regression (*R*) value for the learning data is 0.99, and 0.82 for the validation data and 0.81 for the test data. The overall regression reached 0.81.

Table 6. LSTM regression statistics and MSE.

	Observations	***MSE***	***R***	R^2
Training	16,310	3.1	0.99	0.99
Validation	3495	214.14	0.82	0.67
Test	3495	206.17	0.81	0.66

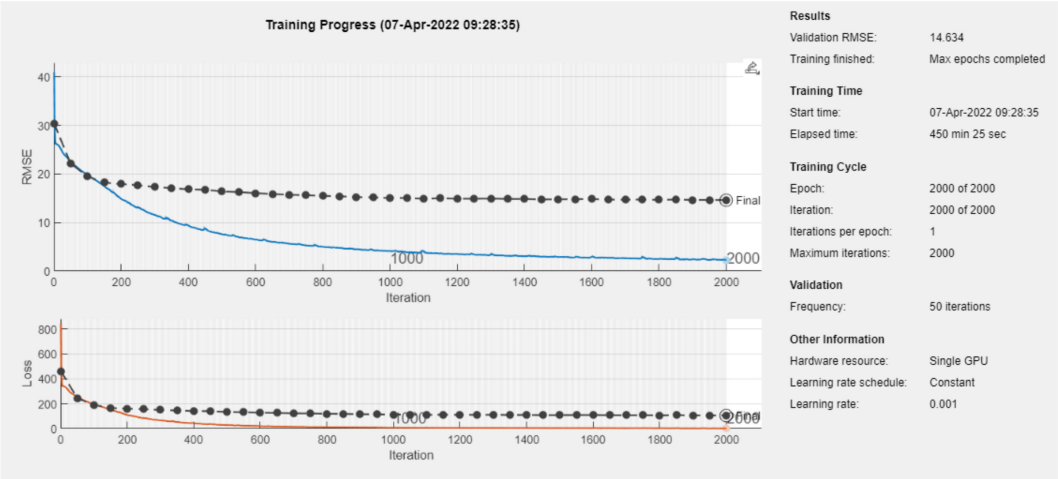


Figure 9. Training progress of the LSTM.

3.4. Selection of the Best Model

While comparing the modeling results of all the methods presented above and evaluating the quality parameters of the developed models, it can be concluded that ANN is the best method in order to predict the level of PM10, as shown in Table 7. Additionally, Table 8 shows the training time and prediction speed for all analyzed models.

Table 7. Quality parameters of all analyzed models.

Quality Parameter	Models Obtained Using Machine Learning Methods					ANN	LSTM
	LR	KNNR	SVM	RT	GPR		
R^2	0.8	0.79	0.82	0.77	0.89	0.90	0.82
MSE	135.51	135.24	119.3	156.57	85.36	68.09	233.52
RMSE	11.64	11.62	10.92	12.51	9.24	8.25	15.28
MAE	8.06	8.02	7.13	8.25	6.12	5.44	9.93

Table 8. Training time and prediction speed for all analyzed models.

	Models Obtained Using Machine Learning Methods					ANN	LSTM
	LR	KNNR	SVM	RT	GPR		
Training time [min]	10:05	00:06	65:11	10:38	129:42	06:49	450:25
Prediction speed [obs/s]	34,000	3380	12,000	78,000	1400	94,000	1500

Taking into account the quality of the presented network for predicting the approximate level of PM10 (R^2 = 0.90, *MSE* = 68.09, *RMSE* = 8.25 and *MAE* = 5.44), it can be concluded that the presented ANN model shows an acceptable level of error, and thus, can be considered a reliable predictor to support decision-making processes. A comparison of real data and those obtained by prediction for this model is shown in Figure 10. Additionally, by analyzing the training time and prediction speed, satisfactory results were also obtained.

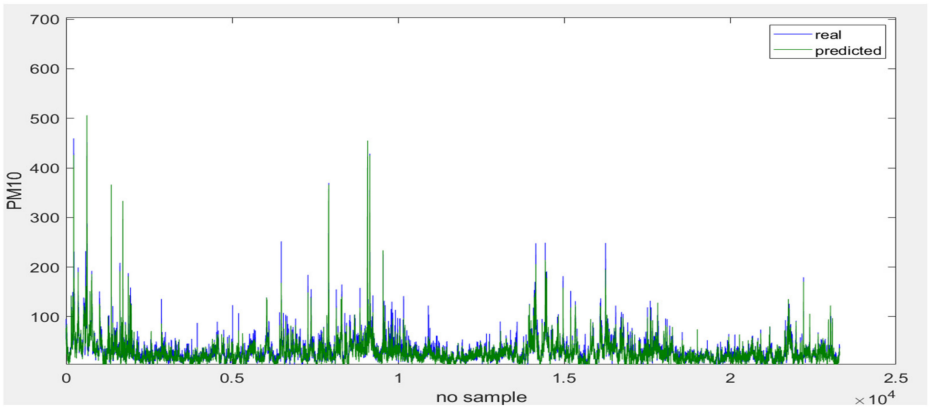


Figure 10. Comparison of the actual data and the data obtained by prediction.

Accordingly, the ANN model was selected for further study. In addition, the ANN network is suitable for such prediction, because one can create models with several outputs and there is no need to create each model separately.

3.5. Prediction Model of Level of PM10 at Different Time Points

The next stage of the study was to use the best model, i.e., ANN to predict the level of PM10 at different time points (PM10 level, PM10 level after 1 h, after 6 h, after 12 h and after 24 h). Input neurons remained unchanged (16 neurons), while the aforementioned five neurons were specified in the output. A schematic representation of the artificial neural network is shown in Figure 11. The number of neurons in the hidden layer (10–700) was

selected experimentally. Other parameters remained unchanged, as in the case of modeling ANN with one neuron at the output.

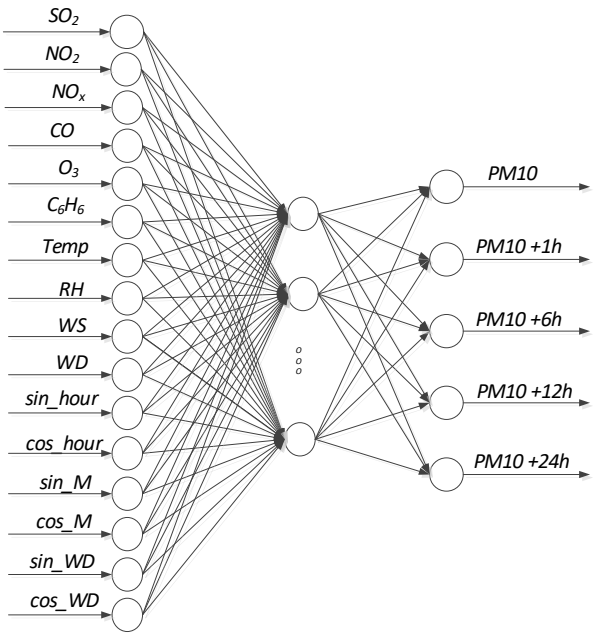


Figure 11. A schematic representation of an artificial neural network.

The best modeling results were obtained for a network with 500 neurons, which was obtained in 28 iterations. The structure of the ANN can be presented as follows: **16** → **500** → **5**, where the first value shows the number of inputs, the second number of neurons in the hidden layer and the third number of output neurons. Other data, such as performance validation, rate of error decrease (gradient) and Mu are presented in Table 9. The best validation performance was obtained for iteration 22, which is shown in Figure 12, while the early stopping strategy is shown in Figure 13.

Table 9. Network training conditions.

Unit	Initial Value	Stopped Value	Target Value
Epoch	0	28	1000
Elapsed Time	-	04:15:51	-
Performance	5.76×10^6	82.8	0
Gradient	8.75×10^6	504	1×10^{-7}
Mu	0.001	0.01	1×10^{10}
Validation Checks	0	6	6

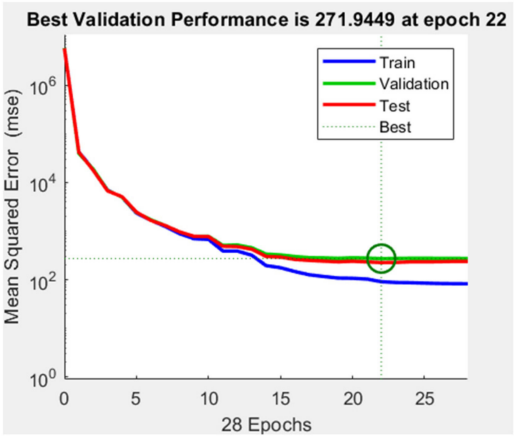


Figure 12. ANN teaching performance for a model with five neurons in the output.

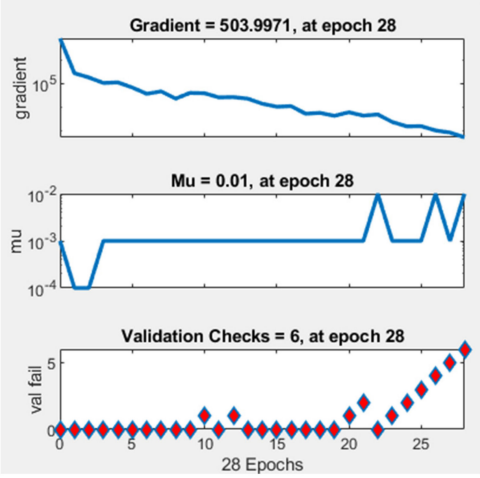


Figure 13. Early stopping strategy for ANN model with five neurons in the output.

Table 10 shows the results of network learning (*MSE* and Regression *R* value) by learning, validation and test subsets. In addition, regression statistics are shown in Figure 14, for which the regression (*R*) value for the learning data is 0.92948, and 0.80 for the validation data and 0.83 for the test data. The overall regression was 0.89383.

Table 10. ANN regression statistics and *MSE*.

	Observations	***MSE***	***R***
Training	16,310	91.24	0.92
Validation	3495	271.94	0.80
Test	3495	225.72	0.83

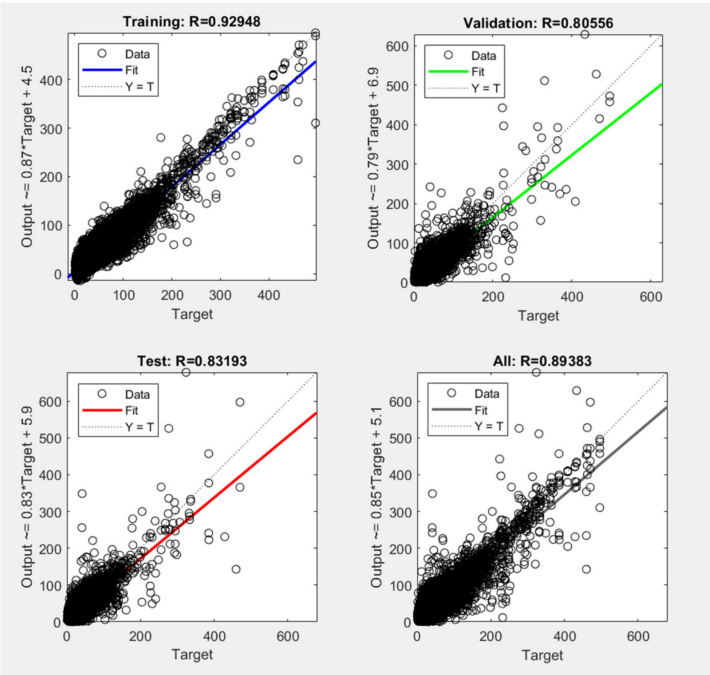


Figure 14. ANN regression statistics for individual sets and the total set.

Considering the quality of the presented network as measured by the level of *MSE* and value of *R* (*R* = 0.89 and *MSE* = 141.89), it can be concluded that the presented ANN model shows an acceptable level of error and can be used to predict the approximate level of PM10. A graphical representation of the prediction obtained by using the ANN model with

five output neurons in comparison with real data at successive time points (PM10 level, PM10 level after 1 h, after 6 h, after 12 h and after 24 h) is shown in Figure 15. Additionally, Figure 15b,d,f,h,j show fragments of these graphs for samples from 23,000 to 23,200.

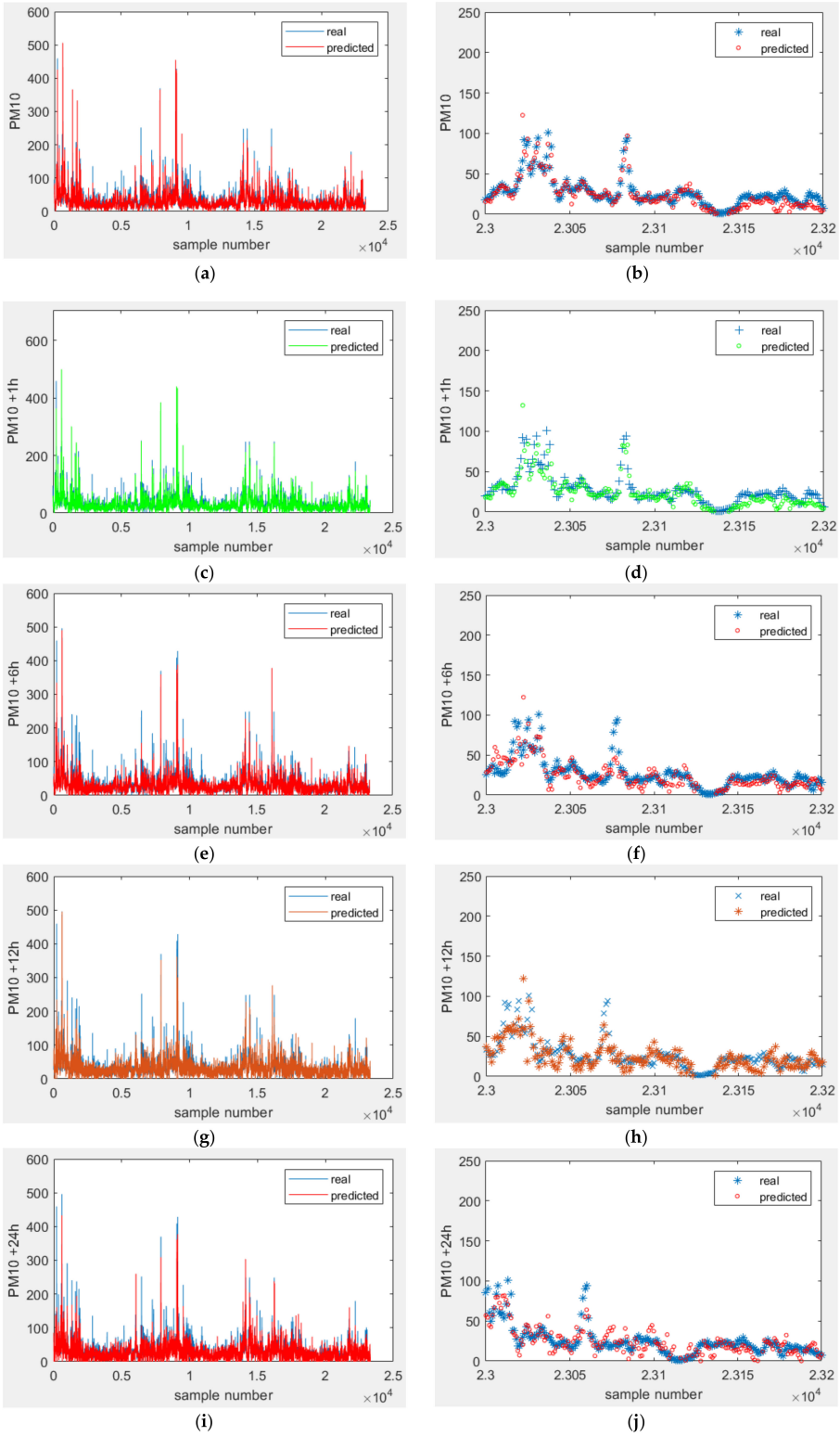


Figure 15. Comparison of actual data and the results obtained by prediction at successive time points: (**a**) PM10 level, (**c**) PM10 level after 1 h, (**e**) PM10 level after 6 h, (**g**) PM10 level after 12 h and (**i**) PM10 level after 24 h; and results for sample numbers ranging from 23,000 to 23,200: (**b**) PM10 level, (**d**) PM10 level after 1 h, (**f**) PM10 level after 6 h, (**h**) PM10 level after 12 h and (**j**) PM10 level after 24 h.

4. Discussion

This paper compares machine learning models for predicting PM10 in the air. Various machine learning algorithms have been used in the literature to predict the air quality parameters. Examples of these algorithms include LASSO regression [36–38], Support Vector Machines (SVM) [39–42], Random Forest [43–46] and k-Nearest Neighbor (kNN) [39,47]. These studies show that machine learning models produce acceptable air pollution forecasting results, can be trained to identify non-linear relationships between input and output data, and are able to predict pollution levels when new data are inserted [48]. The literature shows that the models created in forecasting local levels of particulate matter in the air have a powerful ability to relate complex relationships between input and target variables, directly from the raw data concerning air quality forecasts, and have shown good quality [24,49]. This paper compares machine learning models for predicting PM10 in the air. The models chosen for this study are LR, KNNR, SVM, RT, GPR, ANN and LSTM. Although these models have been partially used for similar purposes, the authors used these methods to predict the amount of PM10 in Lublin using meteorological and chemical air pollution data. One of the novelties of this research is the use of ANN to forecast the amount of PM10 in an area in eastern Poland, which was not previously studied, i.e., Lublin.

In the first phase of the study, the dominant input parameters for the prediction of PM10 involved using LR, KNNR, SVM, RT and GPR. Linear regression was also a tool to investigate the selection of input parameters for ANN and LSTM model. From Table 3, it can be seen that most of the machine learning regression models are characterized by a correlation coefficient $R^2 \geq 0.8$. Only in the case of the GPR model, this coefficient is equal to 0.89, and for this model, it can be assumed to have acceptable quality in predicting the PM10 levels. The lowest *MSE* value was also obtained for this model (*MSE* = 85.36). The linear regression model can be a proper model to predict if the accuracy is sufficient, which is known as one of the simplest machine learning models [50]. As their comparison shows, the best results were obtained for the GPR model. Shahriar et al. (2020) predicted PM10 using the machine learning models, such as linear-support vector machine (L-SVM), medium Gaussian-support vector machine (M-SVM), Gaussian process regression (GPR) and random forest regression (RFR). The modeling results also showed that GPR is the best model for predicting PM10. They used the following variables: NO_X, SO_2, CO and O_3, along with meteorological variables in Dhaka, Chattogram, Rajshahi and Sylhet for the period of 2013 to 2019. Shahriar GPR model achieved the R^2 values ranging from 0.91 to 0.94, while in L-SVM models, R^2 = 0.82–0.89 [51]. A study in Seoul by Jang et al. (2020) used GPR to predict PM10, where the final R^2 value was 0.98. In general, simple Gaussian type models are used for short-range local problems. These models are applicable for pollutant emissions into uniform atmospheric floors [52]. Moreover, they are widely used in regulatory purposes because of their near real time solutions. Unfortunately, these models are not suitable for predicting flow and concentration in complex urban or industrial areas, which are the places where aerosol particles of major concern at present [53].

In the second phase of the study, neural network models were created—shallow neural networks, e.g., ANN, and deep neural networks, e.g., LSTM. Accuracy is crucial in the selection of appropriate input parameters for the development of ANN models, since the accuracy of the created models mainly depends on its structure. ANN has a number of advantages over other traditional modeling approaches, such as handling enormous amounts of data, generalization capabilities, identifying complex relationships between dependent and independent variables and detecting the inherent interactions between process variables [54]. In recent years, some researchers have proposed that long short-term memory (LSTM) networks have higher prediction accuracy. LSTM are probably the most powerful approach to learning from sequential data. The potential of LSTM based models is fully revealed when learning from massive datasets, from which we can detect complex patterns. The LSTM model may have the best predictive ability, but it is greatly affected by the data processing [35,54,55]. Several different algorithms for selecting dependent variables have been presented in the literature: sensitivity analysis, correlation analysis,

multi-objective genetic algorithms and geometric approaches [56]. However, these methods have their advantages and disadvantages, so the selection of output variables is specific and very important for the models created.

In addition, researchers use a variety of input data to predict the amount of PM10 using ANN. Most models are based on meteorological data, others on the data pertaining to chemical pollutants, but there are few that combine both types of data, which is another element of novelty of the research conducted. In the work presented here, the selection of input parameters was made using regression analysis. All the data analyzed in the paper are statistically significant, so they were used for modeling as input parameters: temperature (T), relative humidity (RH), wind speed (WS), wind direction (WD) and air pollution data, including SO_2, PM10, NO_2, NO_x, CO, O_3, C_6H_6, sin_h, cos_h, sin_M cos_M, sin_WD and cos_WD. The results of the parameter quality of the created neural network models show that ANN outperforms the LSTM model, obtaining a higher R^2 value and three times lower *MSE* values. The quality of ANN models is very satisfactory, and therefore, it is worth considering their application in a broader aspect in environmental management.

The next step was to assess the quality of all the analyzed models. The ANN model turned out to the best for predicting the amount of PM10. It is characterized by an R^2 of 0.90 and an *MSE* of 68.09. The *MSE* parameter of the ANN model is half the value that achieved by other models. The R^2 for all machine learning methods is close to 0.8, or close to 1 in the case of ANN. The conclusion from the analysis of the indicators in Table 7 suggests that the machine learning methods have a greater potential for forecasting air pollution without seasonality breakdown.

Other researchers have also succeeded in achieving high R^2 values in machine learning models. Of course, these studies are not directly comparable because they were carried out at different locations with different datasets. Czernecki tested four ML models: AIC-based stepwise regression, two tree-based algorithms (random forests and XGBoost) and neural networks for forecasting PM10 and PM2.5 in four Polish cities (Łódź, Kraków, Poznań and Gdańsk) during the winter season. For both PM10 and PM2.5, the XGBoost algorithm provided the highest correlation values (about 0.98), while the weakest were obtained by AIC (about 0.86) and the ANN R^2 network obtained an R^2 value of 0.919 [32]. Kowalski created the following models: Multiple Linear Regression, Multiple Linear Regression with Regularization and, finally, Linear Neural Networks for PM10 prediction. He obtained the largest R^2 value (R^2 = 0.9256) for Linear Neural Networks [57]. Of course, these studies are not directly comparable, because they were carried out at different locations with different data sets. This study is one of the first to design PM10 machine learning models without seasonality breakdown of input data.

The final step in the analysis presented here was the creation of an ANN model to predict PM10 levels after 1 h, after 6 h, after 12 h and after 24 h with meteorological input data and chemical air pollutants. The model created was characterized by R = 0.89383 and *MSE* = 141.897. The results obtained are not significantly different from those presented in the literature, but there are few models that combine both meteorological as well as chemical inputs and additionally predict PM10 levels at five different time intervals. The use of a single model for simultaneous prediction of PM10 at five different time intervals allows faster determination of the risk of exceeding the PM10 alert level.

The results of the current study are herein compared with the findings of other research studies. This is to serve as the proof that inputting both meteorological and chemical data without seasonality breakdown can work just as well as other available models. While comparing the created model with others (Table 11—values of *R* and *RMSE*), it can be stated that the general goodness of fit between the measured and simulated data is satisfactory. On the basis of the *RMSE* data, it can be surmised that the findings of the current study are an improvement on the status quo. The *RMSE* obtained is lower than those of most other architectures.

The ANN model created for forecasting the PM10 levels and those presented in the literature appear to be a promising tool for air pollution forecasts and could be an alternative to current models.

Table 11. Comparison with other ANN models.

Year, Place	Model	Type of Input Data	Target	*RMSE*	*R*	References
Only winter period (December, January, February) in the period 2002/2003–2016/2021; Gdansk, Gdynia, Sopot, Poland	MLP-ANN	air temperature (AT), relative humidity (RH), air pressure, wind speed (VS)	hourly PM10 concentrations 1–6 h ahead	9.42–23.56	0.50–0.84	[31]
2009–2017, 6 stations in Ankara	ANN	PM10	24-h PM10 concentration	20.8	0.58	[58]
Canetto 2009–2014	ANN	meteorological variables	24-h PM10 concentration	-	0.59	[59]
London, 2007–2012	ANN	Meteorological variables (wind velocities, wind direction, solar radiation, relative humidity, ambient temperature) and the data type (traffic volume, sound level and speeds)	24-h PM10 concentration	-	0.8	[26]
2020, 28 cities of India, 2016–2018	MLP–ANN	PM10, WS, RH, AT, CO_2, NO_2, SO_2, Rainfall, Dew point	PM10 for 1 day ahead	-	0.65	[60]
Kocaeli, Turcja, 120 dni, 2 stacje	ANN	T, RH, AP (hPa), WS direction	PM-10	-	0.74	[61]
Delhi, India, May 2016–May 2018	ANN	PM, CO, SO_2, NOx NO, C_7H_8, NO2, WS, WD (wind direction), VWS (vertical wind speed), RH, Temperature (T), Solar radiation	PM-10	-	0.85	[62]
the model presented in the work	ANN	T, RH, WS, WD, and air pollution data were: SO_2, PM10, NO_2, NO_x, CO, O_3, C_6H_6	PM10 after 1 h, after 6 h, after 12 h and after 24 h	8.25	0.89	

5. Conclusions

This article presents the research on the effectiveness of the use of machine learning methods to predict PM10. The quality of the seven methods was compared. The first of those compared involved machine learning methods: LR, KNNR, SVM, RT and GPR. Linear regression was also a tool to investigate the selection of input parameters for the ANN and LSTM models.

The models were trained on the data from the Lublin-Radawiec meteorological station for 2017–2019; 18 input variables of meteorology and chemical pollutants were used, without considering the seasons. The highest quality was obtained for the ANN model (R^2 = 0.904, *MSE* = 68.09) and the lowest quality for RT (R^2 = 0.77, *MSE* = 156.57).

In addition, an ANN model was created with five output neurons for PM10 prediction after 1 h, after 6 h, after 12 h and after 24 h. The input variables for ANN modeling were selected based on linear regression analysis. The ANN model was characterized by the correlation coefficient of 0.89, as well as *MSE* equal to 91.24, 271.94 and 225.72 for the training, testing and validating set, respectively. These results were obtained with the following set division: training (70%) and testing (30%), using the Neural Network Fitting app, and the Levenberg–Marquardt algorithm for training. The results of the study indicate that ANN models can forecast PM10 best among all the models analyzed, at different time intervals and using the data from meteorological stations.

The ANN approach, therefore, may be useful to effectively derive a predictive understanding of the PM10 concentration level, and thus, provide a tool to the policymakers for improving the decision making associated with air pollution and public health. The created model can be used to predict other pollutants, such as ozone.

In future works, the authors will attempt to classify air pollution levels using machine learning methods.

Author Contributions: Conceptualization, J.K., M.K., P.O. and W.C.; methodology, J.K., M.K., P.O. and W.C.; software, M.K. and P.O.; validation, J.K., M.K., P.O. and W.C.; formal analysis, J.K., M.K., P.O. and W.C.; investigation, J.K. and W.C.; resources, J.K. and W.C.; data curation, J.K., M.K., P.O. and W.C.; writing—original draft preparation, J.K., M.K., P.O. and W.C.; writing—review and editing, J.K., M.K., P.O. and W.C.; visualization, J.K., M.K., P.O. and W.C.; supervision, M.K.; project administration, M.K.; funding acquisition, J.K., M.K., P.O. and W.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Polish Ministry of Science and Higher Education, grant numbers: FD-NZ-020/2022, FD-20/IS-6/019, FD-NZ-030/2022 and FD-20/IS-6/003.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Połednik, B. Emissions of Air Pollution in Industrial and Rural Region in Poland and Health Impacts. *J. Ecol. Eng.* **2022**, *23*, 250–258. [CrossRef]
2. Millán-Martínez, M.; Sánchez-Rodas, D.; Sánchez de la Campa, A.M.; de la Rosa, J. Impact of the SARS-CoV-2 lockdown measures in Southern Spain on PM10 trace element and gaseous pollutant concentrations. *Chemosphere* **2022**, *303*, 134853. [CrossRef] [PubMed]
3. Manisalidis, I.; Stavropoulou, E.; Stavropoulos, A.; Bezirtzoglou, E. Environmental and Health Impacts of Air Pollution: A Review. *Front. Public Health* **2020**, *8*, 14. [CrossRef] [PubMed]
4. Yousaf, H.S.; Abbas, M.; Ghani, N.; Chaudhary, H.; Fatima, A.; Ahmad, Z.; Yasin, S.A. A comparative assessment of air pollutants of smog in wagah border and other sites in Lahore, Pakistan. *Braz. J. Biol.* **2021**, *84*, 1–11. [CrossRef]
5. Regulation of the Minister of Climate and Environment of 11 December 2020 on Assessing the Levels of Substances in the Air (Journal of Laws 2020, item 2279). Available online: https://isap.sejm.gov.pl/isap.nsf/download.xsp/WDU20010620627/U/D20010627Lj.pdf (accessed on 5 March 2022).
6. *PN-EN 12341:2014-07*; Ambient Air—Standard Gravimetric Measurement Method for the Determination of the PM10 or PM2.5 Mass Concentration of Suspended Particulate Matter. European Standards: Brussels, Belgium, 2014.
7. *PN-EN 16450:2017-05*; Ambient Air—Automated Measuring Systems for the Measurement of the Concentration of Particulate Matter (PM10; PM2.5). European Standards: Brussels, Belgium, 2017.
8. Danek, T.; Weglinska, E.; Zareba, M. The influence of meteorological factors and terrain on air pollution concentration and migration: A geostatistical case study from Krakow, Poland. *Sci. Rep.* **2022**, *12*, 11050. [CrossRef]
9. Andrews, B.A. Clean air handbook. *Choice Rev. Online* **2015**, *52*, 52–5100. [CrossRef]
10. Jia, Y.Y.; Wang, Q.; Liu, T. Toxicity research of PM2.5 compositions in vitro. *Int. J. Environ. Res. Public Health* **2017**, *14*, 232. [CrossRef]
11. Sówka, I.; Nych, A.; Kobus, D.; Bezyk, Y.; Zathey, M. Analysis of exposure of inhabitants of Polish cities to air pollution with particulate matters with application of statistical and geostatistical tools. *E3S Web. Conf.* **2019**, *100*, 00075. [CrossRef]
12. Kobus, D.; Merenda, B.; Sówka, I.; Chlebowska-Styś, A.; Wroniszewska, A. Ambient air quality as a condition of effective healthcare therapy on the example of selected polish health resorts. *Atmosphere* **2020**, *11*, 882. [CrossRef]
13. Klimont, Z.; Kupiainen, K.; Heyes, C.; Purohit, P.; Cofala, J.; Rafaj, P.; Borken-Kleefeld, J.; Schöpp, W. Global anthropogenic emissions of particulate matter including black carbon. *Atmos. Chem. Phys.* **2017**, *17*, 8681–8723. [CrossRef]
14. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on Ambient Air Quality and Cleaner Air for Europe. Available online: http://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:32008L0050 (accessed on 5 March 2022).
15. Ordieres, J.B.; Vergara, E.P.; Capuz, R.S.; Salazar, R.E. Neural network prediction model for fine particulate matter (PM 2.5) on the US-Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua). *Environ. Model. Softw.* **2005**, *20*, 547–559. [CrossRef]
16. The National Centre for Emissions Management (KOBiZE). Available online: https://kobize.pl/en/page/id/409/about-us (accessed on 5 March 2022).

17. Chief Inspectorate of Environmental Protection (GIOŚ in Polish) Report on the Forecast of PM2.5 and PM10 Concentrations for 2020 and 2025. 2020. Available online: https://www.lubelskie.pl/file/2020/08/POP_strefa_Aglomeracja_Lubelska_0601.pdf (accessed on 5 March 2022).
18. The Lublin Regional Assembly Air Protection in Lublin Agglomeration. 2020. Available online: https://edziennik.lublin.uw.gov.pl/legalact/2020/4028/ (accessed on 5 March 2022).
19. The Act of 27 April 2001, Environmental Protection Law (Journal of Laws of 2020, item 1219, as Amended). Available online: https://isap.sejm.gov.pl/isap.nsf/download.xsp/WDU20081991227/U/D20081227Lj.pdf (accessed on 5 March 2022).
20. Łobocki, L. Methodological Guidelines for Mathematical Modeling in the Air Quality Management System. Available online: https://www.mos.gov.pl/kategoria/2135_wskazowki_metodyczne_dotyczace_modelowania_matematycznego_w_systemie_zarzadzania_jakoscia_powietrza/ (accessed on 5 March 2022).
21. Institute of Meteorology and Water Management—National Research Institute. Available online: https://imgw.pl/ (accessed on 5 March 2022).
22. Baklanov, A.; Zhang, Y. Advances in air quality modeling and forecasting. *Glob. Transit.* **2020**, *2*, 261–270. [CrossRef]
23. Lu, W.-Z.; Wang, W.-J.; Wang, X.-K.; Yan, S.-H.; Lam, J.C. Potential assessment of a neural network model with PCA/RBF approach for forecasting pollutant trends in Mong Kok urban air, Hong Kong. *Environ. Res.* **2004**, *96*, 79–87. [CrossRef]
24. Azid, A.; Juahir, H.; Toriman, M.E.; Kamarudin, M.K.A.; Saudi, A.S.M.; Hasnam, C.N.C.; Aziz, N.A.A.; Azaman, F.; Latif, M.T.; Zainuddin, S.F.M.; et al. Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. *Water Air Soil Pollut.* **2014**, *225*, 2063. [CrossRef]
25. Arhami, M.; Kamali, N.; Rajabi, M.M. Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations. *Environ. Sci. Pollut. Res.* **2013**, *20*, 4777–4789. [CrossRef]
26. Suleiman, A.; Tight, M.R.; Quinn, A.D. Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM10 and PM2.5). *Atmos. Pollut. Res.* **2019**, *10*, 134–144. [CrossRef]
27. Mehdipour, V.; Stevenson, D.S.; Memarianfard, M.; Sihag, P. Comparing different methods for statistical modeling of particulate matter in Tehran, Iran. *Air Qual. Atmos. Health* **2018**, *11*, 1155–1165. [CrossRef]
28. Krishan, M.; Jha, S.; Das, J.; Singh, A.; Goyal, M.K.; Sekar, C. Air quality modelling using long short-term memory (LSTM) over NCT-Delhi, India. *Air Qual. Atmos. Health* **2019**, *12*, 899–908. [CrossRef]
29. Cai, M.; Yin, Y.; Xie, M. Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach. *Transp. Res. Part D Transp. Environ.* **2009**, *14*, 32–41. [CrossRef]
30. Mao, W.; Wang, W.; Jiao, L.; Zhao, S.; Liu, A. Modeling air quality prediction using a deep learning approach: Method optimization and evaluation. *Sustain. Cities Soc.* **2021**, *65*, 102567. [CrossRef]
31. Nidzgorska-Lencewicz, J. Application of artificial neural networks in the prediction of PM10 levels in thewinter months: A case study in the Tricity Agglomeration, Poland. *Atmosphere* **2018**, *9*, 203. [CrossRef]
32. Czernecki, B.; Marosz, M.; Jędruszkiewicz, J. Assessment of machine learning algorithms in short-term forecasting of pm10 and pm2.5 concentrations in selected polish agglomerations. *Aerosol Air Qual. Res.* **2021**, *21*, 200586. [CrossRef]
33. Matlab R2022a The MathWorks, Inc., Natick, MA, USA. Available online: https://matlab.mathworks.com/ (accessed on 10 March 2022).
34. R 4.1.2 R Foundation for Statistical Computing, Vienna, Austria. Available online: http://www.r-project.org/index.html (accessed on 10 March 2022).
35. Elsheikh, A.H.; Sharshir, S.W.; Elaziz, M.A.; Kabeel, A.E.; Guilan, W.; Haiou, Z. Modeling of solar energy systems using artificial neural network: A comprehensive review. *Sol. Energy* **2021**, *149*, 223–233. [CrossRef]
36. Chu, H.; Wei, J.; Wu, W. Streamflow prediction using LASSO-FCM-DBN approach based on hydro-meteorological condition classification. *J. Hydrol.* **2020**, *580*, 124253. [CrossRef]
37. Son, Y.; Osornio-Vargas, Á.R.; O'Neill, M.S.; Hystad, P.; Texcalac-Sangrador, J.L.; Ohman-Strickland, P.; Meng, Q.; Schwander, S. Land use regression models to assess air pollution exposure in Mexico City using finer spatial and temporal input parameters. *Sci. Total Environ.* **2018**, *639*, 40–48. [CrossRef] [PubMed]
38. Xu, G.; Ren, X.; Xiong, K.; Li, L.; Bi, X.; Wu, Q. Analysis of the driving factors of PM2.5 concentration in the air: A case study of the Yangtze River Delta, China. *Ecol. Indic.* **2020**, *110*, 105889. [CrossRef]
39. Fan, W.; Si, F.; Ren, S.; Yu, C.; Cui, Y.; Wang, P. Integration of continuous restricted Boltzmann machine and SVR in NOx emissions prediction of a tangential firing boiler. *Chemom. Intell. Lab. Syst.* **2019**, *195*, 103870. [CrossRef]
40. Murillo-Escobar, J.; Sepulveda-Suescun, J.P.; Correa, M.A.; Orrego-Metaute, D. Forecasting concentrations of air pollutants using support vector regression improved with particle swarm optimization: Case study in Aburrá Valley, Colombia. *Urban Clim.* **2019**, *29*, 100473. [CrossRef]
41. Saxena, A.; Shekhawat, S. Ambient Air Quality Classification by Grey Wolf Optimizer Based Support Vector Machine. *J. Environ. Public Health* **2017**, *2017*. [CrossRef]
42. Zhu, S.; Lian, X.; Wei, L.; Che, J.; Shen, X.; Yang, L.; Qiu, X.; Liu, X.; Gao, W.; Ren, X.; et al. PM2.5 forecasting using SVR with PSOGSA algorithm based on CEEMD, GRNN and GCA considering meteorological factors. *Atmos. Environ.* **2018**, *183*, 20–32. [CrossRef]
43. Kamińska, J.A. The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: A case study in Wrocław. *J. Environ. Manag.* **2018**, *217*, 164–174. [CrossRef]

44. Rubal; Kumar, D. Evolving Differential evolution method with random forest for prediction of Air Pollution. *Procedia Comput. Sci.* **2018**, *132*, 824–833. [CrossRef]
45. Sun, H.; Gui, D.; Yan, B.; Liu, Y.; Liao, W.; Zhu, Y.; Lu, C.; Zhao, N. Assessing the potential of random forest method for estimating solar radiation using air pollution index. *Energy Convers. Manag.* **2016**, *119*, 121–129. [CrossRef]
46. Wang, Y.; Du, Y.; Wang, J.; Li, T. Calibration of a low-cost PM2.5 monitor using a random forest model. *Environ. Int.* **2019**, *133*, 105161. [CrossRef]
47. Wen, C.; Liu, S.; Yao, X.; Peng, L.; Li, X.; Hu, Y.; Chi, T. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. *Sci. Total Environ.* **2019**, *654*, 1091–1099. [CrossRef]
48. Rahman, M.M.; Shafiullah, M.; Rahman, S.M.; Khondaker, A.N.; Amao, A.; Zahir, M.H. Soft computing applications in air quality modeling: Past, present, and future. *Sustainability* **2020**, *12*, 4045. [CrossRef]
49. Lu, H.C. The statistical characters of PM10 concentration in Taiwan area. *Atmos. Environ.* **2002**, *36*, 491–502. [CrossRef]
50. Kim, M.J.; Yun, J.P.; Yang, J.B.R.; Choi, S.J.; Kim, D. Prediction of the temperature of liquid aluminum and the dissolved hydrogen content in liquid aluminum with a machine learning approach. *Metals* **2020**, *10*, 330. [CrossRef]
51. Shahriar, S.A.; Kayes, I.; Hasan, K.; Salam, M.A.; Chowdhury, S. Applicability of machine learning in modeling of atmospheric particle pollution in Bangladesh. *Air Qual. Atmos. Health* **2020**, *13*, 1247–1256. [CrossRef]
52. Jang, J.; Shin, S.; Lee, H.; Moon, I.C. Forecasting the concentration of particulate matter in the seoul metropolitan area using a gaussian process model. *Sensors* **2020**, *20*, 3845. [CrossRef]
53. Brusca, S.; Famoso, F.; Lanzafame, R.; Mauro, S.; Messina, M.; Strano, S. PM10 Dispersion Modeling by Means of CFD 3D and Eulerian-Lagrangian Models: Analysis and Comparison with Experiments. *Energy Procedia* **2016**, *101*, 329–336. [CrossRef]
54. Elsheikh, A.H.; Saba, A.I.; Elaziz, M.A.; Lu, S.; Shanmugan, S.; Muthuramalingam, T.; Kumar, R.; Mosleh, A.O.; Essa, F.A.; Shehabeldeen, T.A. Deep learning-based forecasting model for COVID-19 outbreak in Saudi Arabia. *Process Saf. Environ. Prot.* **2021**, *149*, 223–233. [CrossRef] [PubMed]
55. Hu, C.; Wu, Q.; Li, H.; Jian, S.; Li, N.; Lou, Z. Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. *Water* **2018**, *10*, 1543. [CrossRef]
56. Voukantsis, D.; Karatzas, K.; Kukkonen, J.; Räsänen, T.; Karppinen, A.; Kolehmainen, M. Intercomparison of air quality data using principal component analysis, and forecasting of PM10 and PM2.5 concentrations using artificial neural networks, in Thessaloniki and Helsinki. *Sci. Total Environ.* **2011**, *409*, 1266–1276. [CrossRef]
57. Kowalski, P.; Warchalowski, W. The comparison of linear models for PM10 and PM2.5 forecasting. *WIT Trans. Ecol. Environ.* **2018**, *230*, 177–187. [CrossRef]
58. Bozdağ, A.; Dokuz, Y.; Gökçek, Ö.B. Spatial prediction of PM10 concentration using machine learning algorithms in Ankara, Turkey. *Environ. Pollut.* **2020**, *263*, 114635. [CrossRef]
59. Tamas, W.; Notton, G.; Paoli, C.; Nivet, M.L.; Voyant, C. Hybridization of air quality forecasting models using machine learning and clustering: An original approach to detect pollutant peaks. *Aerosol Air Qual. Res.* **2016**, *16*, 405–416. [CrossRef]
60. Dutta, A.; Jinsart, W. Air Pollution in Indian Cities and Comparison of MLR, ANN and CART Models for Predicting PM10 Concentrations in Guwahati, India. *Asian J. Atmos. Environ.* **2021**, *15*, 2020131. [CrossRef]
61. Özdemir, U.; Taner, S. Impacts of Meteorological Factors on PM10: Artificial Neural Networks (ANN) and Multiple Linear Regression (MLR) Approaches. *Environ. Forensics* **2014**, *15*, 329–336. [CrossRef]
62. Masood, A.; Ahmad, K. A model for particulate matter (PM2.5) prediction for Delhi based on machine learning approaches. *Procedia Comput. Sci.* **2020**, *167*, 2101–2110. [CrossRef]