Paper No.372

ASSESSMENT OF MACHINE LEARNING ALGORITHMS PERFORMANCE FOR PREDICTING PARTICULATE MATTER USING METEOROLOGICAL AND CHEMICAL VARIABLES: A CASE STUDY OF LAHORE, PAKISTAN

Muhammad Haseeb,
Muhammad Umar Farooq,
Zainab Tahir

ASSESSMENT OF MACHINE LEARNING ALGORITHMS PERFORMANCE FOR PREDICTING PARTICULATE MATTER USING METEOROLOGICAL AND CHEMICAL VARIABLES: A CASE STUDY OF LAHORE, PAKISTAN

By:

Muhammad Haseeb[1], Muhammad Umar Farooq[1], Zainab Tahir[2]

Abstract

Air pollution significantly influences human health, particularly in cities. Globally increased PM2.5 concentrations cause a considerable number of premature deaths each year. Therefore, quantifying and comparing different machine learning (ML) models for predicting PM2.5 concentration is essential. This study aims to use ML (Machine learning) models, linear regression (LR), regression tree (RT), support vector machine (SVM), ensemble regression (ER), Gaussian process regression (GPR), and artificial neural networks (ANN), to predict PM2.5 concentration. The models were trained on six years of data (July 2015–December 2021) by optimizing their hyperparameters. Eight meteorological and chemical indicators were examined, including temperature (T), relative humidity (RH), air pressure (AP), ozone (O3), Sulphur dioxide (SO2), nitrogen dioxide (NO2), Dew point (DP), and Wind speed (WS) as Particulate Matter (PM2.5) predictors. The models' effectiveness was assessed using mean square error (MSE), root mean square error (RMSE), correlation coefficient (R), and coefficient of determination (R2) values. The applied algorithms achieved R2 and RMSE values: 0.72 and 13.52 for LR, 0.8 and 12.156 for RT, 0.82 and 10.28 for SVM, 0.81 and 11.87 for ER, 0.94 and 7.65 for GPR, and for ANN 0.99 and 2.36, respectively. The results suggest that ANN was the best-employed model. Increasing the number of hidden layers in the ANN improved the model's efficiency. The ANN model with three hidden layers showed better results than other models. Using ML models to predict PM2.5 concentration is an important step toward understanding and mitigating the effects of particulate matter on human health and the environment. In the context of Pakistan's persistent air quality challenges, this study's findings are relevant to the overarching theme of "Air Quality

[1] Institute of Environmental Engineering and Research, University of Engineering and Technology Lahore, P.O. Box 54780, Pakistan (Email: mrhaseeb223@gmail.com; umarfarooq@uet.edu.pk).
Correspondence: Dr Muhammad Umar Farooq (umarfarooq@uet.edu.pk)
[2] Department of Space Science, University of the Punjab, 54590, Lahore, Pakistan (Email: zainabtahir244@gmail.com)

Management in Pakistan." As the nation grapples with air pollution's far-reaching consequences on human health and the environment, applying advanced machine learning models to predict PM2.5 concentrations offers a potent tool for proactive mitigation and policy formulation. By embracing these findings, policymakers, researchers, and stakeholders can harness the power of accurate predictions to formulate targeted interventions, optimize resource allocation, and design evidence-based clean air initiatives that align with Pakistan's unique context. This research contributes a crucial step towards the sustainable management of air quality in Pakistan, and paving the way for a healthier, more resilient future for the nation.

Keywords: Air quality; PM concentration prediction; Machine learning; ANN; Lahore; Pakistan.

Highlights:

1. Air pollution's impact on health: PM2.5 linked to premature deaths in cities.

2. Machine learning models compared: Artificial Neural Network (ANN) outperforms others in PM2.5 prediction.

3. ANN model excels: R2 of 0.99, RMSE of 2.36 for PM2.5 prediction.

4. Hidden layers matter: ANN with three layers show superior results.

5. Promising tool for health: Machine Learning (ML) models aids in understanding and mitigating PM2.5 effects.

1. Introduction

Air pollution, one of the most significant threats to human health, is receiving more attention due to this heightened awareness of environmental issues. It's a silent killer claiming many people's lives yearly (Bhuiyan et al. 2021). Over 7 million people worldwide lose their lives annually due to air pollution, with 4.2 million passing away due to outdoor air pollution and 3.8 million deaths due to indoor air pollution (Lei et al. 2022). More factories mean more SO2, PM10, CO2, Ozone (O3), NO2, CO, benzene (C6H6), and other pollutants in the air(Millán-Martínez et al. 2022). Air pollution is rising due to increased industrialization, posing a threat to human health and the environment (Haseeb et al. 2022). This rise in air pollution is largely due to the increased use of automobiles and fossil fuels in private and government facilities (Yousaf et al. 2021). It's important to remember that natural processes like grass pollination, soil erosion, and rock weathering

contribute to air pollution just as much as human activities (Manisalidis et al. 2020).

2. Problem statement

The paper focuses on the issue of air quality management, specifically in the region of Pakistan, where air pollution presents significant risks to both human health and the environment. Significant early mortality is associated with increased levels of PM2.5, a harmful air pollutant. In order to address this matter in a proficient manner, there exists an urgent requirement for precise and streamlined predictive models pertaining to PM2.5 concentrations. Although machine learning (ML) has demonstrated potential in this context, the focus of this research is to assess and compare different ML models, including linear regression, regression tree, support vector machine, ensemble regression, Gaussian process regression, and artificial neural networks, in order to identify the model that provides the most precise predictions. The study aims to offer a comprehensive framework for air quality management in Pakistan, with the objective of enabling policymakers, researchers, and stakeholders to effectively address the challenges posed by particulate matter pollution. By doing so, it seeks to facilitate proactive measures to mitigate the negative impacts of air pollution and develop strategies grounded in empirical evidence, ultimately leading to cleaner air and enhanced public health.

3. Material and Methods

3.1 Study Area Description

Lahore was selected as study area. Lahore has a population of 11.13 million people and a network of forty-two main and busy roadways due to the city's rapid development and population explosion (Khokhar et al. 2016). The chemical, automotive, manufacturing, and pharmaceutical industries are particularly significant in Lahore. Air pollution and related health problems are major problems in this subtropical dry city due to the city's heavy industrial and vehicle traffic (Khan et al. 2016). Winter is particularly prevail by photochemical smog (Jabeen et al. 2021). The average annual temperature in Lahore is 30 °C. With an average high of 37 °C, June is the warmest month of the year. January, an average of 21°C, is often the coolest month in Lahore. A whole 16 °C separates the hottest (June) and coldest (January) months. With an annual average of 838.8 mm, rainfall Lahore is considered to have a semi-dry climate. With an AQI of 161 and PM2.5 and O3 as the most pervasive air pollutants in Lahore, the city has poor air quality according to US standards (Nawaz et al. 2023);(Ali et al.

2022). Several factors, including vehicular and industrial emissions, cause air pollution in Lahore.

3.2 The Dataset

The air pullulation data was collected from two separate departments in Lahore, including the Environmental Protection agency and the US Consulate General, Lahore. The time period covered was from 2015 to 2022. The meteorological department provided the necessary meteorological data. The data pertaining to meteorology consisted of the following: temperature (T), relative humidity (RH), wind speed (WS), dew point (DP), and air pressure (AP), whereas the data pertaining to air pollution comprised of the following: Particulate Matter (PM2.5) Sulphur dioxide (SO2), nitrogen dioxide (NO2), and ozone (O3).

Table-1: Statistical Description of the dataset.

Variables	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
PM2.5 (µg/m3)	4.040	450.839	86.943	57.744	1.987	4.903
NO2 (µg/m3)	3.516	166.98	62.006	30.119	0.378	-0.488
SO2 (µg/m3)	0.14	69.98	30.765	10.181	0.492	0.511
O3 (µg/m3)	1.07	103.04	39.598	19.576	0.439	-0.47
Temperature (oC)	6.9	38.5	24.384	7.37	-0.346	-1.099
Humidity (%)	19.8	95	61.034	14.692	-0.376	-0.108
Air pressure (hPa)	947.606	1026.55	1007.48	8.612	-0.966	3.203
Dew Point (oC)	27.3	-0.1	15.69131	6.544476	-0.03814	-1.2054
Wind Speed (km/h)	28.9	0	8.583971	5.853204	0.710404	-0.13566

3.3 Methods

Linear regression (LR), Support Vector Machine (SVM), Ensemble Regression tree (ERT) Regression Trees (RT), and Gaussian Process Regression (GPR) Models are some of the machine learning techniques that have been tested and validated to predict the particulate matter. The selection of the ANN model's input parameters was also investigated using linear regression. LRs are the most common type of regression model and are used as a first choice. However, because of their limitations for predictions, these are not usually used for forecasting. Models like RTs, SVMs, and GPRs are more flexible and are used for prediction. RTs are easy to understand, can fit and predict quickly, and use low memory. Non-linear support vector machines (SVMs) can produce more accurate

predictions but are more difficult to grasp than their linear counterparts. GPR is distinguished by its high level of accuracy.

An artificial neural network (ANN) is a black-box computational approach that uses interactions between nodes to "circulate" data to neighboring nodes in the network (Yafouz et al. 2022). The ANN models are the most recent application of AI to the problem of predicting air pollution, and they are utilized extensively in both forecasting and prediction. Artificial neural networks (ANNs) are nonlinear models that can handle big datasets with various variables (Zaqoot et al. 2018);(Fogelman et al. 2006);(Jimeno-Sáez et al. 2020). The current study predicted PM levels using ML (Machine learning) algorithms. The ANN model was built in MATLAB Simulink (2021a) with the 'nntool' toolbox. Each layer of the ANN model has a specific purpose, and these layers are as follows: (1) input layer for input parameters (2) hidden layers(inner) with different neurons combination, and (3) output layer. An output value is calculated by each neuron by adding together all of the inputs and weighting them in accordance with the activation function (Dawson et al. 2006); (Abyaneh 2014). In this investigation, the outcome, PM2.5, was predicted using a simple MLP (Multi-layer perceptron) model constructed from three distinct algorithms. Humidity, air pressure, temperature, dew point, wind speed, NO2, SO2, and O3 were used as input factors. Each of the 2 to 20 neutrons in the hidden layers were used for processing the input parameter. To evaluate the efficacy of the ANN model, the number of hidden layers were increased from 1 to 3. The non-linearity of the data was optimized with the help of Tansig and Purelin transfer functions. Additionally, the network was trained using a trio of training algorithms: Bayesian regularization backpropagation (T-BR), and Levenberg-Marquardt backpropagation (T-LM). Numerous iterations were performed to find the optimal configuration of hidden layers neutrons, activation function, and training functions based on ANN architecture. The dataset was divided into two parts: the training set, which comprised 70% of the total, and the test set, which made up the remaining 30%.

3.3.1 Assessment of ANN and other Algorithms

All models were evaluated based on their capacity to forecast PM levels using three different performance indicators: mean squared error (MSE), root mean squared error (RMSE), regression coefficient (R), and coefficient of determination (R2The quality of the created models is indicated by the regression coefficients R and R2 as well as the RMSE. The RMSE should be as low as possible (Elsheikh et al. 2019); (Kujawska et al. 2022).

The following machine learning approaches—LR, SVM, Regression Trees, ER, GPR, and ANN—were tested in the subsequent step: The

table below (Table-2) outlines the parameters that were considered necessary for accurate simulation of these procedures.

Table-2: Modelling parameters based on analyzed methods

Methods	Model Parameters
Linear Regression (LR)	Linear (Robust option off), Interaction Linear (Robust option off), Robust Linear (Robust option on), Stepwise Linear (Max number of sets=1000).
	Hyperparameters functions Disabled
Regression Tree (RT)	While training RT, the minimum leaf size was adjusted. Minimum leaf size—4, 12, and 36—were used in the analysis
	Surrogate decision splits were Off
Support Vector Machine (SVM)	Different Kernel functions were used for training SVM:
	Linear kernel, Quadratic kernel, Cubic kernel. Kernel scale—Automatic; Gaussian kernel (Fine, Medium, coarse). Kernel scale (0.61, 2.4, and 9.8)
	standardize data: true. box constraint, epsilon—Automatic,
Ensemble Regression tree (ERT)	Boosted regression tree and Bagged Regression Tree was trained with hyperparameters option disabled
	Standard leaf size—8, Number of learners—30
Gaussian Process Regression (GPR)	GPR trained using different Kernel functions: Matern 5/2 and Exponential, Rational Quadratic, Squared Exponential,

Methods	Model Parameters
	optimize numeric and standardize parameter: true,
	signal standard deviation and sigma: Automatic
Artificial Neural Networks (ANN)	Two Algorithms were used to train the network:
	The number of neurons in the hidden layers (2–200) was selected experimentally.
	The learning set was 70%, whereas 15% was set for each testing and validation.
	Hidden Layers were tested from 1-3 to check the efficiency of the models

4. Results

Obtained Models Using Machine Learning Methods

Table-3 shows model quality parameters derived from the LR, RT, SVM, ER, and GPR models. Moreover, linear regression was utilized to investigate the selection of ANN input parameters.

Table-3: Performance metrics for machine learning models Quality Parameters

Quality Parameters	Obtained Models During Machine Learning				
	LR	RT	SVM	ER	GPR
R	0.8485	0.8944	0.9055	0.90	O.9695
R2	0.72	0.80	0.82	0.81	0.94
MSE	182.79	147.77	105.59	140.89	58.52
RMSE	13.52	12.156	10.276	11.87	7.65

For Linear Models, the stepwise regression method did not produce a better model. The Linear Model with Interactions algorithm provide the best results, with R2 and RMSE equal to 0.72 and 13.52, respectively.

In the context of algorithms using regression trees the Fine Tree was able to provide the best results when contrasted with the Medium and the Coarse

Trees, which were less accurate. For Regression Tree (RT), the best results were obtained with minimum leaf size-12. In this case, the R2 and RMSE values were equal to 0.80 and 12.156, respectively.

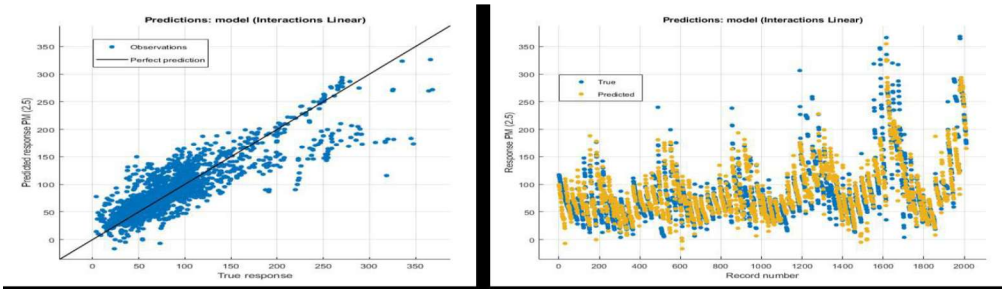
Several computation-focused Kernel functions were tested and compared for support vector machine. There are many different kinds of kernels, such as the Linear kernel, Quadratic kernel, Cubic kernel, and Gaussian or Radial Basis Function (Fine, medium, and coarse) kernel. The training time was shortened by the Fine kernel. The MSE, RMSE, and R2 were also improved using this kernel. For the Support Vector Machine (SVM) the best results were obtained with Fine Gaussian (R2 = 0.82 and RMSE = 10.276).

In case of Ensembled Regression Trees, we employ a pair of learners to create ensemble regression models. 1) Bagged trees, also known as a Bag with decision tree learner, are based on the random forest method. 2) AdaBoost, a boosted tree method using decision trees as the input. The best results were obtained with Bagged tree (R2 = 0.81 and RMSE = 11.87).

A variety of Kernel (Covariance) functions, such as squared exponential, exponential, quadratic, and Matern 52, for use in gaussian process regression (GPR) models were tested. It was discovered that the Gaussian Process Regression Models (GPR) with Rational Quadratic kernel required the least amount of time to train and obtain best results (R2 = 0.94 and RMSE = 7.65)

Models ranking based on RMSE was: GPM –> SVM–>ER–>RT–>LR

Models ranking based on R2 was: GPM –> SVM–>ER–>RT–>LR



(a) **(b)**

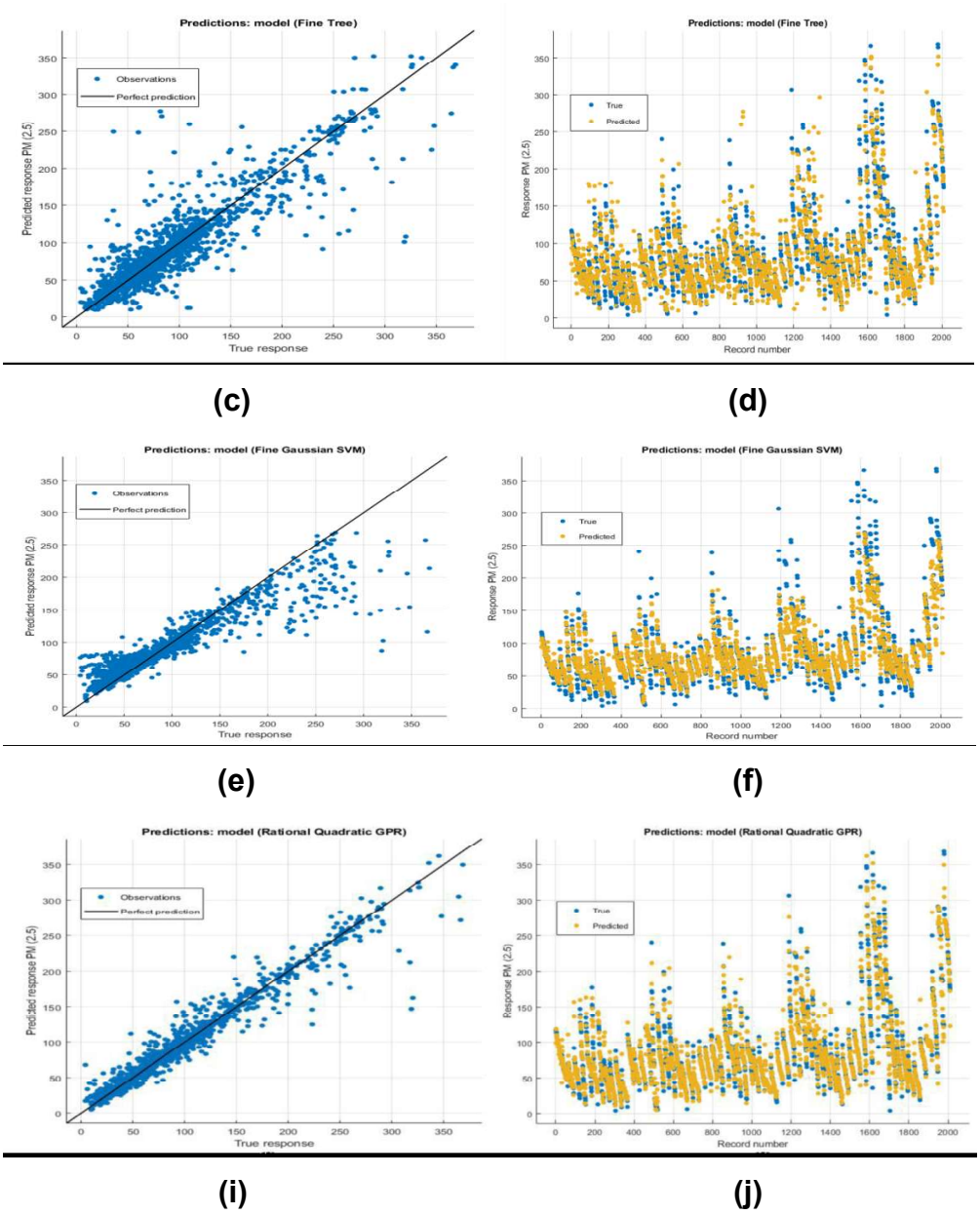


Figure-1: Response plot graphs for the best-selected models (a) LR, (c) RT, (e) SVM, (g) ER, and (i) GPR, and true vs predicted response graphs (b) LR, (d) RT, (f) SVM, (h) ER, and (j) GPR.

4.1 Artificial Neural Network (ANN)

ANN was investigated to make predictions about the PM2.5 concentration. temperature (T), humidity (RH), air Pressure (AP), wind speed (WS), dew

point (DP), SO2, O3, and NO2 were taken as input parameters in the neural network model. There was a sensitivity analysis conducted on T-BR, and T-LM, algorithms with varying numbers of neurons and hidden layers. T-BR algorithm provides the best result with 1,2 and 3 hidden layers combinations with varying neurons while keeping one input and output layer.

Table-4: Best performance model with 1,2 and 3 hidden layers for selected algorithms

Sr. No	Network Type	Training Algorithm	Adoption LF	No of Neurons	MSE	RMSE	R	R2
One Hidden Layer								
1	FFNN	T-LM	GDM	110	98.65	9.9323	0.9730	0.9467
2	FFNN	T-BR	GDM	50	71.29	8.4433	0.9819	0.9641
Two Hidden Layer								
1	FFNN	T-LM	GDM	30-35	58.29	7.6354	0.9812	0.9628
2	FFNN	T-BR	GDM	15-35	10.19	3.1937	0.9930	0.9860
Three Hidden Layer								
1	FFNN	T-LM	GDM	35-40-45	22.26	4.7181	0.9791	0.9586
2	**FFNN**	**T-BR**	**GDM**	**15-17-21**	**5.57**	**2.3601**	**0.9931**	**0.9862**

Bayesian regularization backpropagation (T-BR) achieved the most accurate predictions, with a network consisting of (8-15-17-21-1) three hidden layers. A diagrammatic representation of the ANN model is shown in Figure-2. Table-4 displays additional information, such as performance validation results and error reduction rate for the FFNN ANN models with different parameters. Errors tend to diminish when the number of hidden layers increased, and additional learning iterations was applied to the model. For optimal outcomes, the iteration with the smallest validation error was chosen.

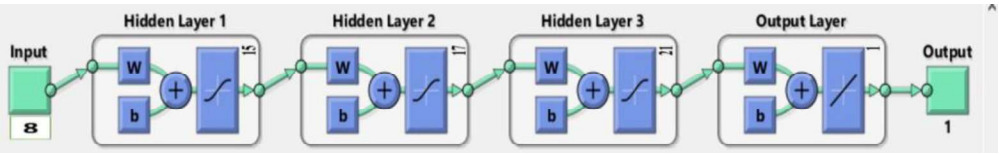


Figure-2: The ANN Structure

The input-hidden-output layer was configured in the same way for every one of these computations (8-'Hidden layers neurons'-1). All of the analyses carried out while ANN was being developed took into consideration and utilized these functions because of the rapid convergence speed and positive output findings they produced

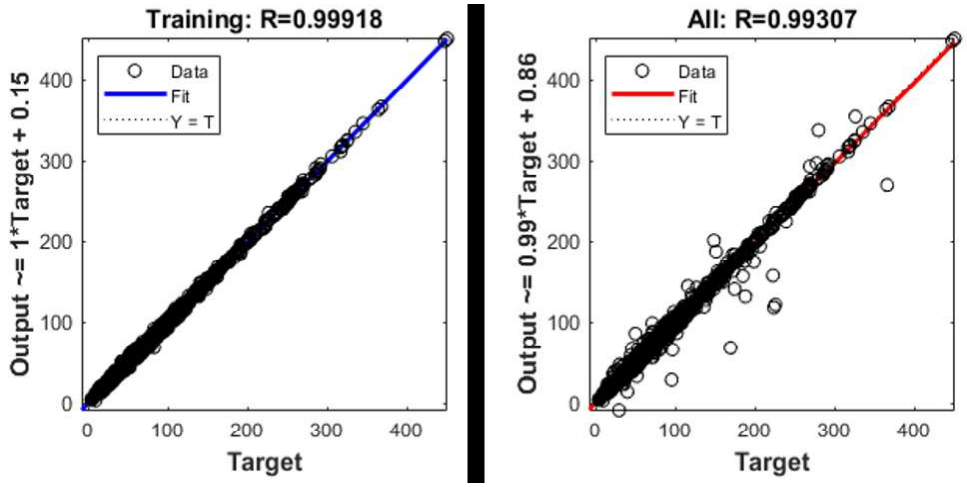


Figure-3: ANN regression statistics for the best trained algorithm (T-BR).

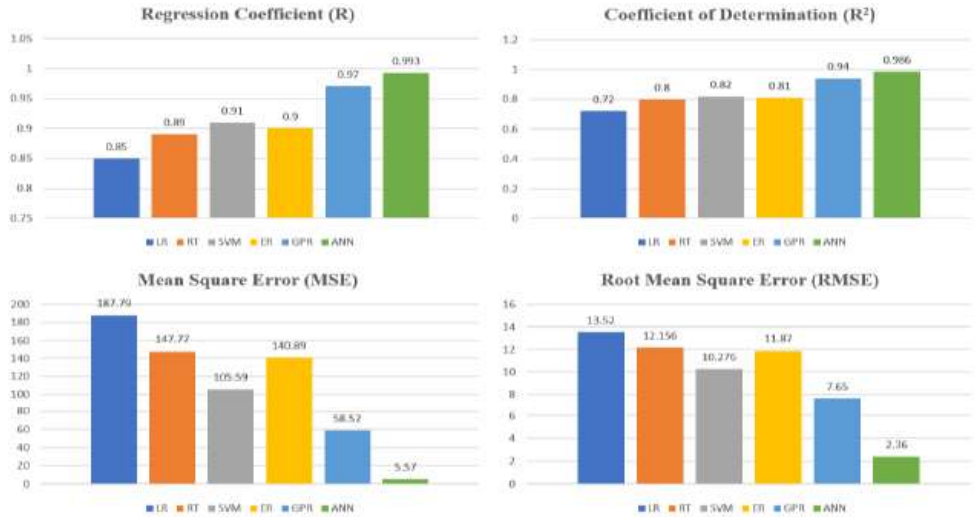


Figure-4: Quality Parameters for all analyzed models

With an R2 of 0.986, MSE of 5.57, and RMSE of 2.36, for predicting of particulate matter concentration, the given ANN model with three hidden Layers has a low enough error rate to be used as a credible predictor in decision-making.

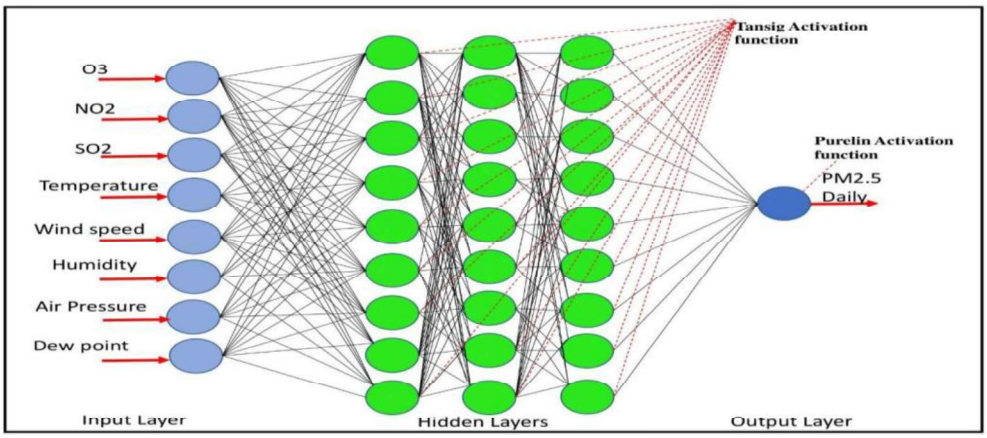


Figure-5: ANN architecture for PM prediction

4.2 PM (2.5) prediction 2015–2021 Trained Data, 2022 Test Data

Based upon the performance, ANN model was used to make predictions about the concentration of PM2.5 for the year 2022. Figure-5 represents a disparity between observed data and predicted data for ANN model with three hidden layers. Results were also satisfactory when considering the training duration and prediction speed**.** The Figure-6 presents both the observed and anticipated levels of PM2.5 concentration in 2022 using the T-BR ANN model. A linear regression analysis was also conducted between actual vs predicted values Figure-7. The model delivered a high R2 of 0.982, which is extremely encouraging for air quality forecasting.

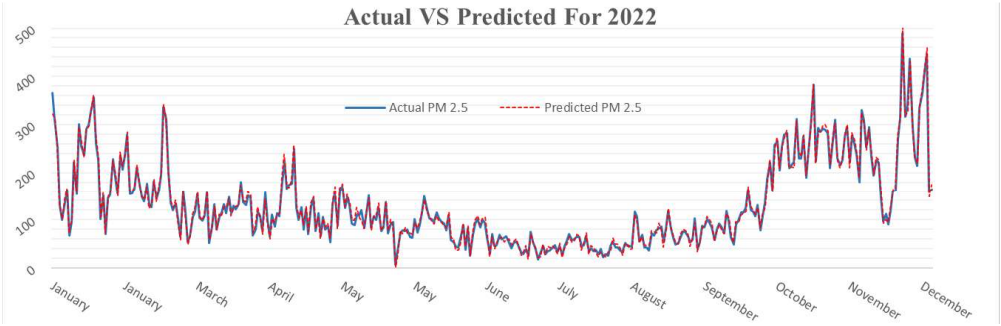


Figure-6: Observed vs predicted PM2.5 concentration (µg/m3) for 2022 using ANN *model*

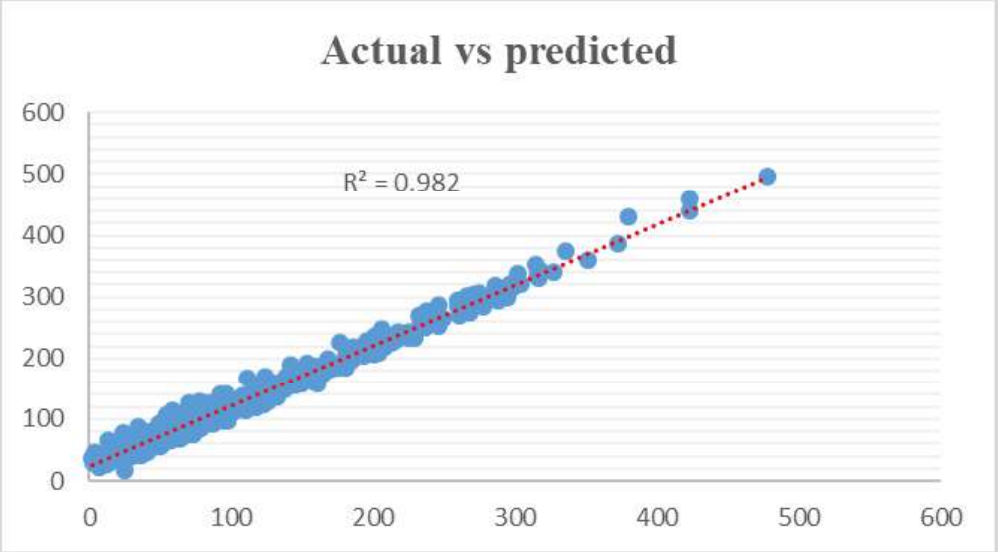


Figure-7: Regression analysis between actual vs predicted value

5. Focus

In this Study, we evaluate and compare various machine learning models for forecasting PM2.5 levels. Parameters of air quality can be predicted using a number of different machine learning techniques, some of which have been discussed in the scientific literature including; k-nearest neighbor (KNN) (Wen et al. 2019);(Fan et al. 2019), Random forest (RF) (Sun et al. 2016);(Kamińska 2018);(Rubal and Kumar 2018);(Wang et al. 2019), long short-term memory (LSTM) (Drewil and Al-Bahadili 2022), Support Vector Machines (SVM) (Saxena and Shekhawat 2017);(Zhu et al. 2018);(Murillo-Escobar et al. 2019);(Fan et al. 2019), and LASSO regression (Son et al. 2018);(Chu et al. 2020);(Xu et al. 2020). Researches demonstrate the effectiveness of machine learning models in forecasting air pollution and their capacity to predict pollution levels even after the introduction of new data (Rahman et al. 2020). According to the research, the models developed for predicting have a strong capacity to link complex interactions between input and target variables, and have shown good quality when applied to raw data about air quality projections (Lu 2002);(Azid et al. 2014).

In this study, various machine learning models were evaluated and compared for the forecasting of atmospheric PM2.5. LR, SVM, RT, ERT, GPR, and ANN were selected as the models to use in this analysis. An innovative aspect of this Study is the incorporation of ANN into forecasting PM2.5 concentration levels in Lahore, a topic that has not before been investigated. Initial analyses employing LR, SVM, RT, ERT, and GPR as the primary machine learning algorithms for PM2.5. Figure-4 shows that the

majority of the machine learning models have an R2> 0.85. In case of GPR model R2 is 0.94, hence only that model can be considered to have adequate quality in predicting particulate matter. This model also achieved the lowest RMSE value (7.65) of any of the others. Although it is one of the simplest machine learning models, if the accuracy is high enough, the it can be used as a suitable for making predictions (Kim et al. 2020). The results showed that with optimized hyperparameters across stations, LR, SVR, GPR, and ANN were able to provide the greatest R2 (83% and 89%) (Yafouz et al. 2022).

ANNs models were built in the second phase of the research. Since the structure of ANN models determines their accuracy. Compared to other modelling techniques, ANN is superior in its ability to process large amounts of data, generalize, find relationships between dependent and independent variables, and detect process variable interactions better than other modelling methods (Elsheikh et al. 2021)..

Table-5: Analysis of Previous literature for the prediction of PM2.5.

Country/place	***Model used***	***Input data***	***Target data***	***RMSE***	***R***	***Reference***
India (Dehli)	ANN	Temperature, Humidity, wind speed, wind direction, solar radiation	PM2.5	-	0.923	(Kaur et al. 2020)
China	ANN and WANN	Precipitation, WD, Air pressure, Temperature, vapor pressure, Humidity, and WS were used as input parameters	PM2.5	-	0.93	(Guo et al. 2023)
China.	RBF and ANN	NOX, Temperature, PM10, CO	PM2.5	7.88	0.92	(Asadollahfardi et al. 2016)
Ahvaz, Iran	ANN	Hourly averaged PM	PM2.5	46.44	0.86	(Goudarzi et al. 2021)
India	ANN	Data from real, urban air quality	PM2.5, NO2, O3	16.4	0.93	(Agarwal et al. 2020)
Ethiopia	ANN	Monitoring network In Dehli traffic volume, weather data and particulate matter concentrations	PM2.5PM10	15.66	-	(Jida et al. 2021)
presented model	ANN	Temperature, Humidity, wind speed, Air pressure, Dew point, NO2, SO2, O3	PM2.5	2.36	0.993	

6. Conclusion

The study evaluates the performance of machine learning techniques for PM2.5 forecasting. Six different models were employed for their relative quality. Machine learning techniques (LR, SVM, RT, ERT, and GPR) were compared. Linear regression analysis was performed for the selection of input parameters for models. Without taking into account of seasonality, the models were trained using Lahore's data from 2015-2021, using 8 input variables of meteorology and air pollutants. With an R2 value of 0.99 and

RMSE of 2.36, the ANN model performed best, whereas the LR model performed lowest (R2 = 0.72 and RMSE = 13.52). The results were achieved by using the neural network fitting, the Bayesian regularization backpropagation (T-BR), and three-hidden-layer architecture to train the network. The training set comprised of 70% of the data while the testing set comprised of 30% data. The study found that, of all the models examined, ANN models performed the best at forecasting PM.

Therefore, the ANN method may be beneficial in efficiently generating a predictive understanding of the PM2.5 concentration, and as a result, it may provide policymakers with a tool to enhance decision making regarding air pollution and public health issues. The created model is capable of being utilized for the purpose of anticipating additional pollutants such as PM.

7. Way Forward

The way forward in this research endeavour involves several crucial steps. First and foremost, further refinement and optimization of the machine learning models, particularly the promising artificial neural networks (ANN) model, should be pursued. This may include exploring advanced neural network architectures and experimenting with different hyperparameter settings to enhance predictive accuracy. Additionally, ongoing data collection and monitoring of PM2.5 and relevant meteorological and chemical indicators are essential to ensure the models' continued relevance and accuracy. Collaboration with local environmental and health authorities in Pakistan is vital for the successful implementation of predictive models into real-time air quality management systems. Furthermore, this research lays the foundation for proactive policy formulation and targeted interventions to mitigate air pollution's impact on public health and the environment. Continued research and data-driven strategies will be instrumental in fostering a cleaner and healthier future for Pakistan, aligning with the broader goal of sustainable air quality management in the region.

8. Reference:

Abdullah S, Napi NNLM, Ahmed AN, et al (2020) Development of multiple linear regression for particulate matter (PM10) forecasting during episodic transboundary haze event in Malaysia. Atmosphere 11:1–14. https://doi.org/10.3390/atmos11030289

Abyaneh HZ (2014) Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. Journal of Environmental Health Science and Engineering 12:6–13. https://doi.org/10.1186/2052-336X-12-40

Agarwal S, Sharma S, R. S, et al (2020) Air quality forecasting using artificial neural networks with real time dynamic error correction in highly polluted regions. Science of the Total Environment 735:139454. https://doi.org/10.1016/j.scitotenv.2020.139454

Ali M, Siddique I, Abbas S (2022) Characterizing Air Pollution and Its Association with Emission Sources in Lahore: A Guide to Adaptation Action Plans to Control Pollution and Smog. Applied Sciences 12:5102. https://doi.org/10.3390/app12105102

Ameer S, Shah MA, Khan A, et al (2019) Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities. IEEE Access 7:128325–128338. https://doi.org/10.1109/ACCESS.2019.2925082

Amini H, Trang Nhung NT, Schindler C, et al (2019) Short-term associations between daily mortality and ambient particulate matter, nitrogen dioxide, and the air quality index in a Middle Eastern megacity. Environmental Pollution 254:113121. https://doi.org/https://doi.org/10.1016/j.envpol.2019.113121

Asadollahfardi G, Zangooei H, Aria SH (2016) Predicting PM2.5 Concentrations Using Artificial Neural Networks and Markov Chain, a Case Study Karaj City. Asian Journal of Atmospheric Environment 10:67–79. https://doi.org/10.5572/ajae.2016.10.2.067

Awang NR, Elbayoumi M, Ramli NA, Yahaya AS (2016) Diurnal variations of ground-level ozone in three port cities in Malaysia. Air Quality, Atmosphere & Health 9:25–39. https://doi.org/10.1007/s11869-015-0334-7

Azid A, Juahir H, Toriman ME, et al (2014) Prediction of the Level of Air Pollution Using Principal Component Analysis and Artificial Neural Network Techniques: a Case Study in Malaysia. Water, Air, & Soil Pollution 225:2063. https://doi.org/10.1007/s11270-014-2063-1

Bekesiene S, Meidute-Kavaliauskiene I, Vasiliauskiene V (2021) Accurate prediction of concentration changes in ozone as an air pollutant by multiple linear regression and artificial neural networks. Mathematics 9:1–21. https://doi.org/10.3390/math9040356

Bhuiyan MAM, Sahi RK, Islam MR, Mahmud S (2021) Machine learning techniques applied to predict tropospheric ozone in a semi-arid climate region. Mathematics 9:1–13. https://doi.org/10.3390/math9222901

Cai M, Yin Y, Xie M (2009) Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach. Transportation

Research Part D: Transport and Environment 14:32–41. https://doi.org/10.1016/j.trd.2008.10.004

Castro M, Pires JCM (2019) Decision support tool to improve the spatial distribution of air quality monitoring sites. Atmospheric Pollution Research 10:827–834. https://doi.org/10.1016/j.apr.2018.12.011

Chu H, Wei J, Wu W (2020) Streamflow prediction using LASSO-FCM-DBN approach based on hydro-meteorological condition classification. Journal of Hydrology 580:124253. https://doi.org/10.1016/j.jhydrol.2019.124253

Dashti Latif S, Najah Ahmed A, Sherif M, et al (2021) Reservoir water balance simulation model utilizing machine learning algorithm. Alexandria Engineering Journal 60:1365–1378. https://doi.org/10.1016/j.aej.2020.10.057

Dawson CW, Abrahart RJ, Shamseldin AY, Wilby RL (2006) Flood estimation at ungauged sites using artificial neural networks. Journal of Hydrology 319:391–409. https://doi.org/10.1016/j.jhydrol.2005.07.032

Doreswamy, Harishkumar KS, Km Y, Gad I (2020) Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models. Procedia Computer Science 171:2057–2066. https://doi.org/10.1016/j.procs.2020.04.221

Drewil GI, Al-Bahadili RJ (2022) Air pollution prediction using LSTM deep learning and metaheuristics algorithms. Measurement: Sensors 24:100546. https://doi.org/10.1016/j.measen.2022.100546

Elsheikh AH, Saba AI, Elaziz MA, et al (2021) Deep learning-based forecasting model for COVID-19 outbreak in Saudi Arabia. Process Safety and Environmental Protection 149:223–233. https://doi.org/https://doi.org/10.1016/j.psep.2020.10.048

Elsheikh AH, Sharshir SW, Abd Elaziz M, et al (2019) Modeling of solar energy systems using artificial neural network: A comprehensive review. Solar Energy 180:622–639. https://doi.org/10.1016/j.solener.2019.01.037

Fan W, Si F, Ren S, et al (2019) Integration of continuous restricted Boltzmann machine and SVR in NOx emissions prediction of a tangential firing boiler. Chemometrics and Intelligent Laboratory Systems 195:. https://doi.org/10.1016/j.chemolab.2019.103870

Fogelman S, Zhao H, Blumenstein M, Zhang S (2006) Estimation of oxygen demand levels using UV-Vis spectroscopy and artificial neural networks as an effective tool for real-time, wastewater treatment control

Galvão SLJ, Matos JCO, Kitagawa YKL, et al (2022) Particulate Matter Forecasting Using Different Deep Neural Network Topologies and Wavelets for Feature Augmentation. Atmosphere 13:1451. https://doi.org/10.3390/atmos13091451

Goudarzi G, Hopke PK, Yazdani M (2021) Forecasting PM2.5 concentration using artificial neural network and its health effects in Ahvaz, Iran. Chemosphere 283:131285. https://doi.org/https://doi.org/10.1016/j.chemosphere.2021.131285

Guo Q, He Z, Wang Z (2023) Predicting of Daily PM2.5 Concentration Employing Wavelet Artificial Neural Networks Based on Meteorological Elements in Shanghai, China. Toxics 11:51. https://doi.org/10.3390/toxics11010051

Haseeb M, Tahir Z, Batool SA, et al (2022) The Carbon Footprint of a Public Sector University Before and During the COVID-19 Lockdown. Global NEST Journal 24:29–36. https://doi.org/10.30955/gnj.004222

Hashim NIM, Noor NM, Annas S (2018) Influence of meteorological factors on variations of particulate matter (PM10) concentration during haze episodes in Malaysia. AIP Conference Proceedings 2045:. https://doi.org/10.1063/1.5080916

Ibrahim RK, Afan HA, El-shafie A, Fai CM (2019) Dual Scenarios

Ismail M, Abdullah S, Jaafar AD, et al (2018) Statistical modeling approaches for PM10 forecasting at industrial areas of Malaysia. AIP Conference Proceedings 2020:. https://doi.org/10.1063/1.5062670

Jabeen F, Ali Z, Maharjan A (2021) Assessing health impacts of winter smog in lahore for exposed occupational groups. Atmosphere 12:1–14. https://doi.org/10.3390/atmos12111532

Jida SN, Hetet J-F, Chesse P, Guadie A (2021) Roadside vehicle particulate matter concentration estimation using artificial neural network model in Addis Ababa, Ethiopia. Journal of Environmental Sciences 101:428–439. https://doi.org/https://doi.org/10.1016/j.jes.2020.08.018

Jimeno-Sáez P, Senent-Aparicio J, Cecilia JM, Pérez-Sánchez J (2020) Using machine-learning algorithms for eutrophication modeling: Case study of mar menor lagoon (spain). International Journal of Environmental Research and Public Health 17:. https://doi.org/10.3390/ijerph17041189

Kamińska JA (2018) The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: A case

study in Wrocław. Journal of Environmental Management 217:164–174. https://doi.org/10.1016/j.jenvman.2018.03.094

Kaur M, Arvind P, Mandal A (2020) ANN-Based Prediction of PM2.5 for Delhi. In: Ahmed S, Abbas SM, Zia H (eds) Smart Cities---Opportunities and Challenges. Springer Singapore, Singapore, pp 633–644

Khan MM, Zaman K, Irfan D, et al (2016) Triangular relationship among energy consumption, air pollution and water resources in Pakistan. Journal of Cleaner Production 112:1375–1385. https://doi.org/10.1016/j.jclepro.2015.01.094

Khokhar MF, Mehdi H, Abbas Z, Javed Z (2016) Temporal assessment of NO2 pollution levels in urban centers of Pakistan by employing ground-based and satellite observations. Aerosol and Air Quality Research 16:1854–1867. https://doi.org/10.4209/aaqr.2015.08.0518

Kim MJ, Yun JP, Yang JBR, et al (2020) Prediction of the temperature of liquid aluminum and the dissolved hydrogen content in liquid aluminum with a machine learning approach. Metals 10:. https://doi.org/10.3390/met10030330

Krishan M, Jha S, Das J, et al (2019) Air quality modelling using long short-term memory (LSTM) over NCT-Delhi, India. Air Quality, Atmosphere & Health 12:899–908. https://doi.org/10.1007/s11869-019-00696-7

Kujawska J, Kulisz M, Oleszczuk P, Cel W (2022) Machine Learning Methods to Forecast the Concentration of PM10 in Lublin, Poland. Energies 15:1–23. https://doi.org/10.3390/en15176428

Kumari KM (2019) 24 Hour Advance Forecast of Surface Ozone Using Linear and Non-Linear Models at a Semi-Urban Site of Indo-Gangetic Plain. International Journal of Environmental Sciences & Natural Resources 18:. https://doi.org/10.19080/ijesnr.2019.18.555982

Lei TMT, Siu SWI, Monjardino J, et al (2022) Using Machine Learning Methods to Forecast Air Quality: A Case Study in Macao. Atmosphere 13:1–14. https://doi.org/10.3390/atmos13091412

Lu HC (2002) The statistical characters of PM10 concentration in Taiwan area. Atmospheric Environment 36:491–502. https://doi.org/10.1016/S1352-2310(01)00245-X

Manisalidis I, Stavropoulou E, Stavropoulos A, Bezirtzoglou E (2020) Environmental and Health Impacts of Air Pollution: A Review. Frontiers in Public Health 8:. https://doi.org/10.3389/fpubh.2020.00014

Mao W, Wang W, Jiao L, et al (2021) Modeling air quality prediction using a deep learning approach: Method optimization and evaluation. Sustainable Cities and Society 65:102567. https://doi.org/10.1016/j.scs.2020.102567

Millán-Martínez M, Sánchez-Rodas D, Sánchez de la Campa AM, de la Rosa J (2022) Impact of the SARS-CoV-2 lockdown measures in Southern Spain on PM10 trace element and gaseous pollutant concentrations. Chemosphere 303:. https://doi.org/10.1016/j.chemosphere.2022.134853

Murillo-Escobar J, Sepulveda-Suescun JP, Correa M, Orrego-Metaute D (2019) Forecasting concentrations of air pollutants using support vector regression improved with particle swarm optimization: Case study in Aburrá Valley, Colombia. Urban Climate 29:100473. https://doi.org/10.1016/j.uclim.2019.100473

Nawaz R, Aslam M, Nasim I, et al (2023) Air Pollution Tolerance Index and Heavy Metals Accumulation of Tree Species for Sustainable Environmental Management in Megacity of Lahore. 55–68

Occupational WHO, Team EH (2006) WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide : global update 2005 : summary of risk assessment. WHO/SDE/PHE/OEH/06.02

Osman A, Afan HA, Allawi MF, et al (2020) Adaptive Fast Orthogonal Search (FOS) algorithm for forecasting streamflow. Journal of Hydrology 586:. https://doi.org/10.1016/j.jhydrol.2020.124896

Othman F, Alaaeldin ME, Seyam M, et al (2020) Efficient river water quality index prediction considering minimal number of inputs variables. Engineering Applications of Computational Fluid Mechanics 14:751–763. https://doi.org/10.1080/19942060.2020.1760942

Pawlak I, Jarosławski J (2019) Forecasting of surface ozone concentration by using artificial neural networks in rural and urban areas in central Poland. Atmosphere 10:. https://doi.org/10.3390/atmos10020052

Perceptron-colliding UM, Latif SD, Afan HA, Banadkooki FB (2020) Pipeline Scour Rates Prediction-Based Model

Petelin D, Mlakar P, Božnar MZ, et al (2015) Ozone forecasting using an online updating Gaussian-process model. International Journal of Environment and Pollution 57:111–122. https://doi.org/10.1504/IJEP.2015.074494

Rahman MM, Shafiullah M, Rahman SM, et al (2020) Soft computing applications in air quality modeling: Past, present, and future. Sustainability (Switzerland) 12:1–33. https://doi.org/10.3390/SU12104045

ubal, Kumar D (2018) Evolving Differential evolution method with random forest for prediction of Air Pollution. Procedia Computer Science 132:824–833. https://doi.org/10.1016/j.procs.2018.05.094

Sapitang M, Ridwan WM, Kushiar KF, et al (2020) Machine learning application in reservoir water level forecasting for sustainable hydropower generation strategy. Sustainability (Switzerland) 12:. https://doi.org/10.3390/su12156121

Saxena A, Shekhawat S (2017) Ambient Air Quality Classification by Grey Wolf Optimizer Based Support Vector Machine. Journal of Environmental and Public Health 2017:. https://doi.org/10.1155/2017/3131083

Son Y, Osornio-Vargas ÁR, O'Neill MS, et al (2018) Land use regression models to assess air pollution exposure in Mexico City using finer spatial and temporal input parameters. Science of the Total Environment 639:40–48. https://doi.org/10.1016/j.scitotenv.2018.05.144

Sun H, Gui D, Yan B, et al (2016) Assessing the potential of random forest method for estimating solar radiation using air pollution index. Energy Conversion and Management 119:121–129. https://doi.org/10.1016/j.enconman.2016.04.051

Taşpınar F (2015) Improving artificial neural network model predictions of daily average PM10 concentrations by applying principle component analysis and implementing seasonal models. Journal of the Air and Waste Management Association 65:800–809. https://doi.org/10.1080/10962247.2015.1019652

Thupeng WM, Mothupi T, Mokgweetsi B, ... (2018) A Principal Component Regression Model, for Forecasting Daily Peak Ambient Ground Level Ozone Concentrations, in the Presence of Multicollinearity Amongst …. … Journal of Applied … 7:1–12

Tikhamarine Y, Doudja S-G, Najah A-M, et al (2019) Improving Artificial Intelligence Models Accuracy for Monthly Streamflow Forecasting Using Grey Wolf Optimization (GWO) Algorithm. Journal of Hydrology 582:124435. https://doi.org/10.1016/j.jhydrol.2019.124435

Tikhamarine Y, Souag-Gamane D, Najah Ahmed A, et al (2020) Improving artificial intelligence models accuracy for monthly streamflow forecasting using grey Wolf optimization (GWO) algorithm. Journal of Hydrology 582:124435. https://doi.org/10.1016/j.jhydrol.2019.124435

Ul-Saufie AZ, Yahaya AS, Ramli NA, et al (2013) Future daily PM10 concentrations prediction by combining regression models and feedforward

backpropagation models with principle component analysis (PCA). Atmospheric Environment 77:621–630. https://doi.org/10.1016/j.atmosenv.2013.05.017

Wang Y, Du Y, Wang J, Li T (2019) Calibration of a low-cost PM2.5 monitor using a random forest model. Environment International 133:105161. https://doi.org/10.1016/j.envint.2019.105161

Wen C, Liu S, Yao X, et al (2019) A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. Science of the Total Environment 654:1091–1099. https://doi.org/10.1016/j.scitotenv.2018.11.086

Xu G, Ren X, Xiong K, et al (2020) Analysis of the driving factors of PM2.5 concentration in the air: A case study of the Yangtze River Delta, China. Ecological Indicators 110:105889. https://doi.org/10.1016/j.ecolind.2019.105889

Yafouz A, Ahmed AN, Zaini N, El-Shafie A (2021) Ozone Concentration Forecasting Based on Artificial Intelligence Techniques: A Systematic Review. Water, Air, and Soil Pollution 232:. https://doi.org/10.1007/s11270-021-04989-5

Yafouz A, AlDahoul N, Birima AH, et al (2022) Comprehensive comparison of various machine learning algorithms for short-term ozone concentration prediction. Alexandria Engineering Journal 61:4607–4622. https://doi.org/https://doi.org/10.1016/j.aej.2021.10.021

Yan YH, Chen T Bin, Yang CP, et al (2022) Long-term exposure to particulate matter was associated with increased dementia risk using both traditional approaches and novel machine learning methods. Scientific Reports 12:1–9. https://doi.org/10.1038/s41598-022-22100-8

Yousaf HS, Abbas M, Ghani N, et al (2021) A comparative assessment of air pollutants of smog in wagah border and other sites in Lahore, Pakistan. Brazilian journal of biology = Revista brasleira de biologia 84:e252471. https://doi.org/10.1590/1519-6984.252471

Zaqoot HA, Hamada M, Miqdad S (2018) A Comparative Study Of Ann For Predicting Nitrate Concentration In Groundwater Wells In The Southern Area Of Gaza Strip. Applied Artificial Intelligence 32:727–744. https://doi.org/10.1080/08839514.2018.1506970

Zhang YF, Fitch P, Thorburn PJ (2020) Predicting the trend of dissolved oxygen based on the kPCA-RNN model. Water (Switzerland) 12:1–15. https://doi.org/10.3390/w12020585

Zhu S, Lian X, Wei L, et al (2018) PM2.5 forecasting using SVR with PSOGSA algorithm based on CEEMD, GRNN and GCA considering meteorological factors. Atmospheric Environment 183:20–32. https://doi.org/10.1016/j.atmosenv.2018.04.004