World Academy of Science, Engineering and Technology
International Journal of Environmental and Ecological Engineering
Vol:18, No:7, 2024

A Comparative Analysis of Machine Learning Techniques for PM10 Forecasting in Vilnius

M. A. S. Fahim, J. Sužiedelytė Visockienė

Open Science Index, Environmental and Ecological Engineering Vol:18, No:7, 2024 publications.waset.org/10013700.pdf

***Abstract*—**With the growing concern over air pollution (AP), it is clear that this has gained more prominence than ever before. The level of consciousness has increased and a sense of knowledge now has to be forwarded as a duty by those enlightened enough to disseminate it to others. This realization often comes after an understanding of how poor air quality indices (AQI) damage human health. The study focuses on assessing air pollution prediction models specifically for Lithuania, addressing a substantial need for empirical research within the region. Concentrating on Vilnius, it specifically examines particulate matter concentrations 10 micrometers or less in diameter (PM10). Utilizing Gaussian Process Regression (GPR) and Regression Tree Ensemble, and Regression Tree methodologies, predictive forecasting models are validated and tested using hourly data from January 2020 to December 2022. The study explores the classification of AP data into anthropogenic and natural sources, the impact of AP on human health, and its connection to cardiovascular diseases. The study revealed varying levels of accuracy among the models, with GPR achieving the highest accuracy, indicated by an RMSE of 4.14 in validation and 3.89 in testing.

***Keywords*—**Air pollution, anthropogenic and natural sources, machine learning, Gaussian process regression, tree ensemble, forecasting models, particulate matter.

I. Introduction

AIR pollution (AP) seriously threatens human health. It is not only the root of numerous respiratory diseases but it can also generate acute cardiovascular and neurological complications. Since AP has serious implications for public health, it ought to be an important issue to address and predict the levels of AP with precision. Several studies have demonstrated the relevance of AP and conducted the use of forecasting methods. It also compels the decision-makers to enact appropriate measures to curb and manage the human-induced contributors to AP. There is unanimous agreement on the valuable role that the predictive and forecasting models play in this field, especially in helping the efforts to mitigate AP.

Many research studies have delved into the issue at hand aiming to grasp and forecast air dynamics. They have utilized forecasting techniques to show the patterns of air pollution. This joint effort highlights a move towards utilizing models not solely for academic purposes but as crucial instruments for policymakers and environmental authorities. These models could serve as a link between data-driven understandings and practical interventions providing optimism, in the endeavors to reduce human-caused air pollution sources. Focusing on examinations concerning air pollution predictions, this study shows that pollutant PM10 concentrations are influenced by meteorological conditions like wind speed and dew point on atmospheric PM10 concentrations within the Vilnius region (Lithuania). By using techniques of Gaussian Process Regression (GPR), Regression Tree Ensemble, and Regression Tree methodologies, predictive forecasting models have been validated, and tested. These models used time and meteorological parameters as variables to predict PM10 concentrations, utilizing hourly data from January 2020 to December 2022.

Mina Adel Shokry Fahim is with the Vilnius Gediminas Technical University, Lithuania (e-mail: mina-adel-shokry.fahim@vilniustech.lt).

II. Literature Review

Various research investigations have looked into predicting air pollution showing the importance of addressing air quality issues. These studies utilize prediction models and methods for air pollutants and the environmental domain. For instance, using Neural Networks (NN), [1] focused on the prediction of PM2.5 pollution related to health issues and deemed carcinogenic by the World Health Organization (WHO). In the study, deep learning techniques, including long-short term memory units (LSTM), recurrent neural networks (RNN), and gated recurrent unit (GRU) were used to predict the level of PM2.5 in Turkey, based on 1-3 hours of measurements from data from the AP quality monitoring stations in the Ankara District. In this manner, the highest accuracy in forecasting was achieved by an ensemble of GRU and RNN models, determining results of 0.832, 0.709, and 0.611 for one, two, and three-hour forecasts, respectively. Another interesting case involves predictions for many pollutants, such as sulfur dioxide (SO_2), carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), PM2.5 and PM10, using multiple linear regression (MLR), decision tree, random forest (RF) algorithms and others [2]. The research concludes that, among the algorithms tested, RF provides the best prediction for poor AQI. The model depicts an R^2 of 0.79 for O_3, CO, and PM10, 0.701 for NO_2, and 0.86 for PM2.5.

Especially in advanced economies and countries such as Taiwan, between 2012 and 2017, [3] used the dataset of air quality monitoring in Taiwan to make a prediction for PM2.5 levels using ML algorithms. By using the improvements to the conventional methods utilized with Gradient Boosting (XGBoost) as a model for PM2.5 level forecasting achieved an impressive R^2 value.

The air pollution index (API) is a comprehensive indicator capturing pollutants like PM2.5, PM10, CO, SO_2, NO_2, and O_3, among others. Research was conducted in Malaysia, employing

Open Science Index, Environmental and Ecological Engineering Vol:18, No:7, 2024 publications.waset.org/10013700.pdf

the Support Vector Machine (SVM) for its predictive tasks and leveraging its regression capabilities to derive API values from certain inputs. To optimize the model, techniques were applied to handle missing data and outliers, with the removal of outliers notably improving the model's accuracy. The study also revealed that different configurations of SVM parameters produced varied outcomes. Ultimately, the radial basis function (RBF), Kernel function, was selected for the best-result model [4] indicating a great prediction performance. Other investigation based on the MLR model, for forecasting long-term PM10 levels during different monsoon periods, was conducted in Kuala Terengganu, Malaysia and emphasized the opportunity to assess the PM10 concentrations which serve as a major environmental and health threat. The study observed significant impacts during the Northeast Monsoon (NEM) period, while the lowest occurred during the Southwest Monsoon (SWM), with data observed from 2005-2011. Relative humidity, temperature, precipitation, atmospheric pressure, and wind speed were deemed to be critical weather elements that affect PM10 levels. Humidity and precipitation showed an inverse relationship with PM10 levels, scored R^2 ranging from 0.570 to 0.681, providing valuable insights into the seasonal impacts on air quality [5].

NN, k-nearest neighbors, and decision trees are three commonly used machine learning algorithms. Another research worked on a dataset of AP (Access Points) data from Macedonia in 2016. As a result, the algorithm NN achieved the highest accuracy. Despite the fact that the decision trees have the advantage of faster processing of information and ease of understanding, they turned out to be less accurate than the k-nearest neighbor and NN algorithms [6], of 92.3%. This accuracy was verified using confusion matrices. In the United States: Los Angeles, Houston, and Atlanta [7]. Intended to evaluate the performance of three different ML algorithms: SVM, RF, and RNN. The focus was on comparing these algorithms and measuring their effectiveness. The results of this research showed that the RNN model actually has significant accuracy, compared to the other algorithms. Classified AP levels in Atlanta, Houston, and Los Angeles (USA) by using RNNs and SVM. Accuracies of 80.27%, 76.71%, and 76.44% were achieved with the highest accuracy for RNN for Atlanta city. Kulkarni and other researchers presented a system that consistently monitored air quality, providing AQI forecasts for up to 15 hours. Through the application of the SVM algorithm, the study endeavored to predict AQI over the specified period. The model measured levels of pollutants such as PM2.5, PM10, NO, NO_2, and NH_3. Data from the Indian air quality database was divided, with 70% used for model training and 30% for testing. The accuracy of the SVM approach was ascertained using the RMSE metric, which showed a score of 13.25 [8].

A study by [9] introduced an application designed to predict AP using various algorithms. They utilized historical data from Valencia, Spain, spanning two years (2013 and 2014). This data was gathered every hour from three monitoring stations: Moli, Francia, and Silla. The study accounted for traffic intensity and meteorological factors and the findings indicated that the RF algorithm outperformed others in many instances. The mean squared error (MSE) for the prediction models across the three stations' measurements. It is noteworthy that the RF algorithm demonstrated superior results for most tests across the individual stations. In Jordan, where a variety of machine learning algorithms, including the Multi-Layer Perceptron Neural Network (MLP), Decision Tree Regression (DTR), and XGBoost, were utilized for forecasting O_3 concentration levels for the next day. The dataset was assessed using filters such as Holt-Winters smoothing and Savitzky-Golay. It was observed that the MLP algorithm was superior, and the incorporation of the Savitzky-Golay filter significantly enhanced the prediction accuracy, improving the R^2 by 50% and reducing both the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) by 80% [10].

Combining Convolutional Neural Networks (CNN), LSTM networks, and GPR into a CLSTM-GPR which have been done by [11] The model accurately forecast PM2.5 concentrations. By effectively harnessing spatial-temporal data, the model surpasses existing CNN-GPR, LSTM-GPR, and GPR models in both point prediction accuracy and interval reliability. Demonstrated at two monitoring stations, the CLSTM-GPR model was the best performance, results with R increasing by over 4.38%, R^2 increasing by over 8.96%, MAE decreasing by over 5.14%, RMSE decreasing by over 4.68%.

III. Sources and Impacts of Air Pollution

AP data is classified into two categories: anthropogenic and natural sources. Anthropogenic sources include stationary factories, waste incinerators, and mobile sources like cars, ships, and planes. Other contributions include controlled burning practices in agriculture and forest management, mining activities, fumes from aerosols and solvents, landfill waste disposal (producing methane), and military activity such as nuclear testing [12].

On the other hand, natural sources of pollution data include dust from barren areas, methane from animal digestion, radon from radioactive decay, CO from wildfires, and compounds like sulfur, chlorine, and ash from volcanic eruptions [13].

Anthropogenic AP affects human health. Researching the quality of the atmosphere is essential because of its ties to well-being complications. Unwanted elements in the sky can cause wellness issues, underlining an international matter of concern [14]. End-of-life occurrences are significant. Heart-related ailments (cardiovascular disease (CVD)) top the list of causes of global mortality. Roughly 17.9 million individuals succumb to CVD annually, making up 31% of total fatalities. In Malaysia, CVD is the primary cause of death, accounting for 35% of all deaths [15]. There is a strong connection between air quality issues and CVD, with 60-80% of deaths related to air contaminants being attributed to CVD [15], [16]. Diverse elements lead to AP, stemming from a blend of several harmful components. These encompass airborne particles and specific gases, such as O_3, CO, and NO_2. Both human actions and natural phenomena are the culprits. While human-related emissions can be managed, natural occurrences, atmospheric characteristics encompass elements that outline the atmospheric composition. Elements like wind, its course, and moisture

Open Science Index, Environmental and Ecological Engineering Vol:18, No:7, 2024 publications.waset.org/10013700.pdf

levels significantly influence how pollutants spread and accumulate. Moreover, a rise in atmospheric warmth can shift these contaminants. It is noteworthy that, while the atmosphere can influence pollution levels, pollution, conversely, impacts our climate and urban structure. The layout and structure of a city play a pivotal role in how pollution disperses. Key structural elements include the layout of the streets, the number of inhabitants, land utilization, and contaminant variety. The nature of the contaminant, whether it be gaseous or particulate, holds significance. The density, velocity, and length of time that vehicles are on the road in each area can lead to varying pollutant distribution levels [17].

One of the paramount factors of significant importance is the AQI, a metric employed by governmental and environmental agencies. It delineates a spectrum denoting the degree of cleanliness and healthiness of the air. A higher numerical value within this index signifies a deteriorating condition of AP, conversely reflecting better air quality when the value is lower. Whenever AQI values are between 0 and 50, it comes under the 'Excellent' category, which means it serves a low risk. When AQI falls between 51 and 100, it is 'Good' and air quality is satisfactory, with very little risk. The status 'Lightly Polluted' corresponds to AQI values between 101 and 150, which is an acceptable air quality for those who are sensitive. AQI values between 151 and 200 are classified as 'Moderately Polluted,' indicating that it begins to have an effect on health, especially for sensitive groups. Any AQI values between 201 and 300 fall under the 'Heavily Polluted' range pertaining to health emergency conditions with significant impacts. Values exceeding 300 are categorized as 'Severely Polluted', which refer to extreme health risk conditions for individuals [18]. Many countries have a system that publish daily AQI values as shown in Fig. 1. For instance, the Vilnius Air Quality Index (2023).

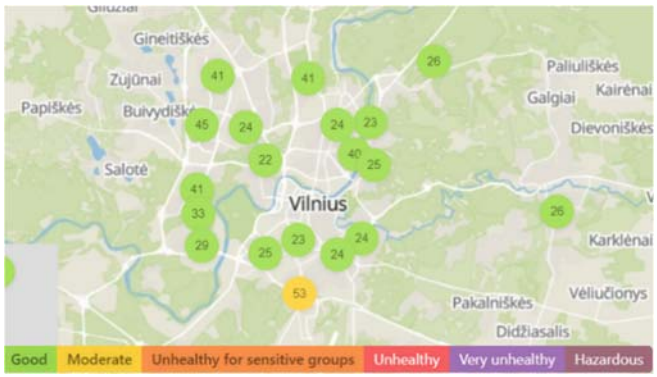

Fig. 1 The air quality in Vilnius City on 2023-10-20 day [19]

IV. Air Pollution Assessment Algorithms

Modern pollution prediction tools are essential, leveraging data-driven approaches to anticipate pollution patterns. Their main benefit is enabling preventive actions, resulting in purer city environments, and supporting choices that encourage sustainable growth. Traditional ways to evaluate quality are often inexact and require long mathematical calculations. Being a part of Artificial Intelligence, ML is an attractive solution. Several of these studies highlighted the capacity of ML to forecast air conditions, as detailed in Fig. 2. The general steps for using ML in the process of prediction focuses its potential to benefit preciseness in AQI forecasts [20]. The algorithms explored in our investigation, based on the literature review, include GPR, Regression Tree Ensemble, and Regression Tree.

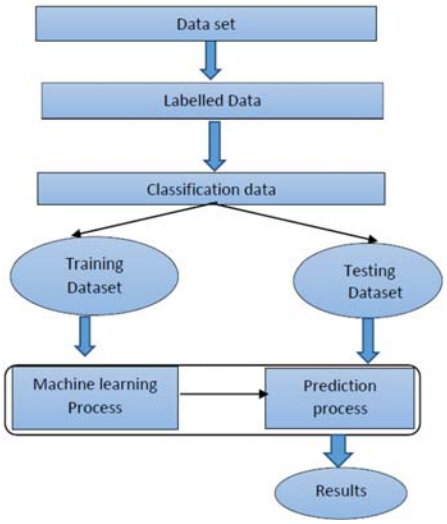

Fig. 2 General steps for using ML Algorithms [20]

The decision tree algorithm is part of the supervised algorithms. It is versatile and can address both regression and classification challenges. The process happens in a series of steps until a decision is reached. RF is an ML technique that combines decision trees to make predictions. During training, it constructs multiple decision trees. Then, it determines the most likely class (for classification tasks) or average prediction (for regression tasks) for new unseen data. By aggregating the results from decision trees, RF effectively reduces overfitting issues and delivers higher accuracy compared to using a single decision tree. In other words, it is like creating a 'forest' of decision trees and then utilizing their combined insights to arrive at a prediction, through voting.

XGBoost is flexible and suitable for both regression and classification problems, since it utilizes gradient-boosted decision trees. However, it is more efficient, can generate decision trees in parallel and is designed to maximize hardware optimization through caching. Moreover, it is feasible to deploy XGBoost on a computer cluster [3].

NN is an ML algorithm concept that is close to the biological neural networks (BNN) in the human brain. BNN are complex networks of neurons and their connections in our brains are responsible for information processing, as well as transmitting. An algorithm was been developed to reproduce these networks, to a certain extent and many studies have been conducted in this area [21].

RNN are a type of NN that can use their past outputs as current inputs, possessing internal memory in the form of hidden layers [22]. LSTM is a type of RNN architecture. They were introduced by [23] to address the limitations of traditional

Open Science Index, Environmental and Ecological Engineering Vol:18, No:7, 2024 publications.waset.org/10013700.pdf

RNNs, particularly the vanishing and exploding gradient problems. These issues made it difficult for standard RNNs to learn and remember long sequences and led to poor performance in tasks that required learning from long-term dependencies. GRU is a sort of RNN architecture proposed by [24]. It was developed to overcome some of the problems associated with traditional RNNs, especially the vanishing gradient problem, similar to LSTM Networks. However, GRUs involve a simplified gating mechanism compared to LSTMs and are more computationally efficient under some circumstances.

MLP is another artificial NN architecture; it consists of multiple layers: an input layer, several hidden layers, and an output layer. Each layer contains nodes or neurons, which are interconnected by weighted pathways. Basically, the intuition behind MLP is its ability to learn and model the non-linear relationships in data. When input is fed into an MLP, it propagates through these layers and gets transformed. It processes data by applying a weighted sum to the input from every neuron and then it goes through an activation function, which is usually non-linear in nature rather than linear (sigmoid or ReLU). The final output is produced after the input has passed through all the layers. The strength of MLPs lies in their capacity for approximating complex functions. To train an MLP, one adjusts the weights of the connections to reduce the discrepancy between its predictions and the actual target values. This training process typically employs back-propagation, a method where the error is propagated backward through the network [10].

SVM is a supervised learning algorithm, predominantly utilized for classification and regression tasks. It stands as an efficient classifier tailored for linear problem-solving. And to address non-linear scenarios, SVM incorporates kernel functions [25]. The main idea behind SVM is to find a hyperplane that best separates the data into classes. In the case of a two-class problem, the optimal hyperplane is the one that has the maximum margin between two classes. Data points that are closest to the hyperplane and influence its position and orientation are called 'support vectors'. Another popular kernel function that helps in learning the non-linear relations of data points in ML is the RBF kernel, also known as a Gaussian kernel. It functions on the basis of distance between two data points and keeps the value lessening as more variance is achieved for feature space. This means that the RBF kernel weights the closer data points with larger values, while those which are farther apart have smaller values [26]. It indirectly maps the input data to a space with higher dimensions; therefore, the data can be linearly separable. As in (1) for the RBF kernel, when comparing two data points (x1 and x2).

$$\mathrm{K}(\mathrm{x1},\mathrm{x2}) = \exp\ (-\gamma[\![\ \|\ \mathrm{x1}-\mathrm{x2}\ \|\ ]\!]\hat{}2\ ) \qquad (1)$$

where γ is a parameter that determines the spread or shape of the kernel.

MLR techniques have been employed in the area of predictive analytics for some time. For instance, MLR has worked well to model a relationship between two or three variables by fitting the given variables into a linear forecasting (2), as suggested in [5].

$$y = b_0 + \sum_{i=1}^{n} b_i . x_i + \varepsilon. \qquad (2)$$

where b_i are the regression coefficients and ε is the stochastic error associated with the regression.

Quantile Regression (QR) is a statistical method designed to determine conditional quantile functions. As opposed to linear regression, which estimates conditional mean functions by minimizing the sum of the squared differences, QR estimates conditional median functions by minimizing unevenly weighted absolute differences. Using quantiles has the primary advantage of providing a more detailed and accurate relationship between random variables than by simply using mean regression [27].

V. Materials and Methods

A. Data Preparation

In our research, we have employed Vilnius, Lithuania as the focal study area. We used and compiled an hourly dataset, which, after went thorough cleaning process, resulted in approximately 22,867 records for PM10, also quantified in μg/m³ spanning from January 2020 to December 2022. As depicted in Fig. 3, this dataset primarily focuses on air pollution metrics, sourced from the European Environment Agency (EEA) [28]. These measurements were taken from the National Air Monitoring Network specifically at the "Vilnius - Senamiestis" air quality station, which is located at coordinates 25.2852 longitude and 54.6776 latitude. In addition to air quality data, the study also incorporates meteorological data, including wind speed, temperature, and dew point. This data was collected from the " Vilnius International, LH" station, provided by the National Center for Environmental Information [29].

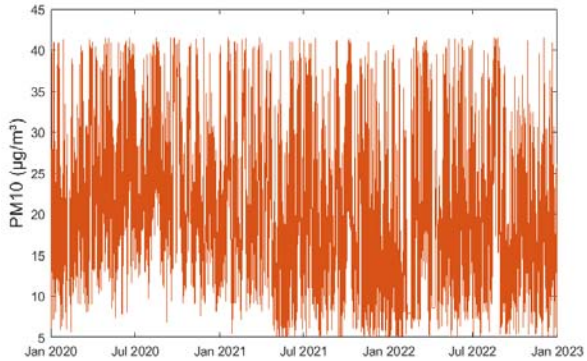

Fig. 3 Timeline with PM10 concentrations

In the applied methodology for data preparation, data from the years 2020 to 2022 were used; 10% was set aside for testing, while 25% was allocated for validation. The dataset underwent a cleaning process, where irrelevant variables such as station identifiers were removed to streamline the dataset. Subsequently, the datasets were merged into a single file for a

Open Science Index, Environmental and Ecological Engineering Vol:18, No:7, 2024 publications.waset.org/10013700.pdf

consolidated analysis. The wind speed, temperature, and dew point were extracted from complex strings and converted into numerical values for analytical purposes all the calculations have been made using MATLAB. The data have been filtered out regarding to verification condition for each observation. Outliers identified through the interquartile range method were removed to ensure data integrity. Finally, the cleaned data, with standardized timestamps and relevant variables, the characteristics of PM10 pollution and meteorological conditions are summarized in Table I. This data preparation methodology ensures the reliability and accuracy of the analysis for insightful research findings.

TABLE I
STATISTICAL SUMMARY OF SEASONAL PM10 VALUES AND METEOROLOGICAL CONDITIONS

Variable	PM10 (µg/m³)	TMP (°C)	DEW (°C)	WND Speed (km/h)
Mean	19.77	7.85	3.94	37.9
Median	18.71	8	4	36
Standard Deviation	7.8	8.84	7.75	19.24
Minimum	5	-18	-20	0
Maximum	41.63	31	22	88
Range	36.63	49	42	88

B. Methods

GPR has become a powerful method for forecasting concentrations of particulate matter (PM), taking advantage of low number of needed parameters, and flexibility in managing intricate challenges like limited sample sizes, nonlinear relationships, and complex dimensional spaces. Esteemed for its dependability and extensive usage in numerous academic researches, the study utilized a regression learner application, incorporating Principal Component Analysis (PCA) with an emphasis on achieving a specified explained variance of 95%. This approach was chosen to reduce the dimensionality of the dataset. The nonparametric characteristics of GPR ensure that any limited collection of random variables can be efficiently modeled with a multivariate Gaussian distribution. This feature is critical for addressing complex regression issues and sorting initial data distributions. Additionally, the wide range of covariance functions within GPR facilitates the development of various continuous structures. In [30], the rational quadratic covariance function, as mentioned in (3) within the regression model, was selected for its ability to handle varied scales of data and to model a more extensive range of functions compared to more straightforward kernels.

$$k(x, x') = \left(1 + \frac{|x - x'|^2}{2\alpha l^2}\right)^{-\alpha} \quad (3)$$

where x, x' are two points in the input space, $|x - x'|^2$ is the squared Euclidean distance between x and x', l^2 is the length scale of the kernel, which determines the "smoothness" of the resulting function, α is a scale mixture parameter that controls the weight of large- and small-scale variations in the data [31].

A Regression Tree is a decision tree designed for continuous outcome prediction. It segments the predictor space into distinct regions, using a series of binary decisions optimized to reduce variance within each region. Predictions are made based on the mean outcome of the training observations within each leaf node [32]. Bagged Regression Trees, or Bootstrap Aggregating, improve prediction robustness and accuracy by training multiple regression trees on bootstrapped dataset samples and averaging their predictions. This technique reduces variance without significantly increasing bias, detailing its effectiveness in mitigating overfitting while enhancing predictive performance in regression tasks.

C. Accuracy Indicators

The indicator accuracy of models comes from the calculation of RMSE, MAE, MSE, and the R^2. RMSE is extremely useful as a number to indicate a model's performance, in the case of training, cross-validation stage, or monitoring the model performance after deployment. The formula is as follows:

$$\text{RSME} = \sqrt{\frac{\sum(y_i - \widehat{y_i})^2}{N-P}} \quad (4)$$

where y_i is the actual value for the i^{th} observation, $\hat{y}_i$ is the predicted value for the i^{th} observation, N is the number of observations, and P is the number of parameter estimates, including the constant.

The RMSE is used to evaluate the regression algorithms. In order to define the MAE, the absolute difference of the measurements for a device under analysis, from the measurements of the reference instrument, is calculated. It represents the average of the absolute value of deviations for the two values. MAE presented in (5) is expressed in the same units as the measurements and ranges from 0 to infinity. A low MAE means that the measurements from a device under testing are very close, in absolute value, to the measurements from the reference instrument. A high MAE means that measurements are far from those of the reference instrument, in terms of absolute value [33].

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad (5)$$

where $\hat{y}_i$ is the prediction value, x_i is the true value, and n is the total number of data points.

It is advised to compute R^2 as in (6) in addition to MAE.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (6)$$

y_i is the actual observation value, $\hat{y}_i$ is the model prediction value, $\bar{y}$ is the sample mean value, and n is the number of samples.

In analyzing the data from experiments, it should be noted that the MAE metric is not a very good indicator of calibrated correlated deviations, or deviations which are the result of random errors or limitations of the device under analysis. In most cases, they cannot be corrected. All kinds of deviations increase MAE, equally. For example, if an analyzed device achieves a high R^2 due to perfect tracking with changes in the concentration of pollutants but is not quite accurate because of

poor calibration, then its MAE might not be dissimilar from that of a broken device that always outputs the same number. A high R^2 can point to the possibility of the to calibrate the device under analysis [34].

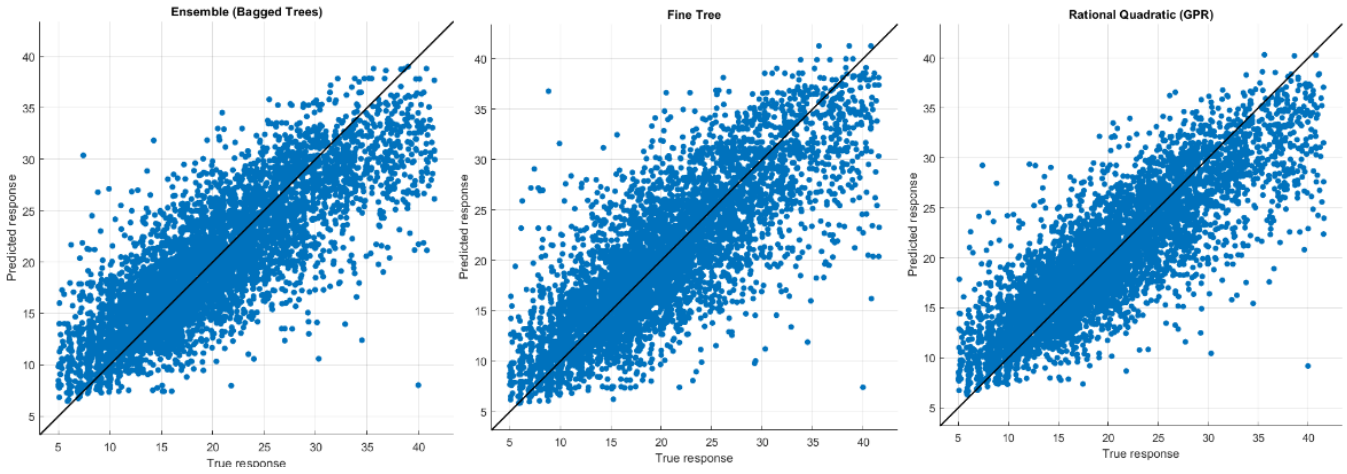

Fig. 4 Scatter plots of predicted response of PM10 concentrations for the validation phase

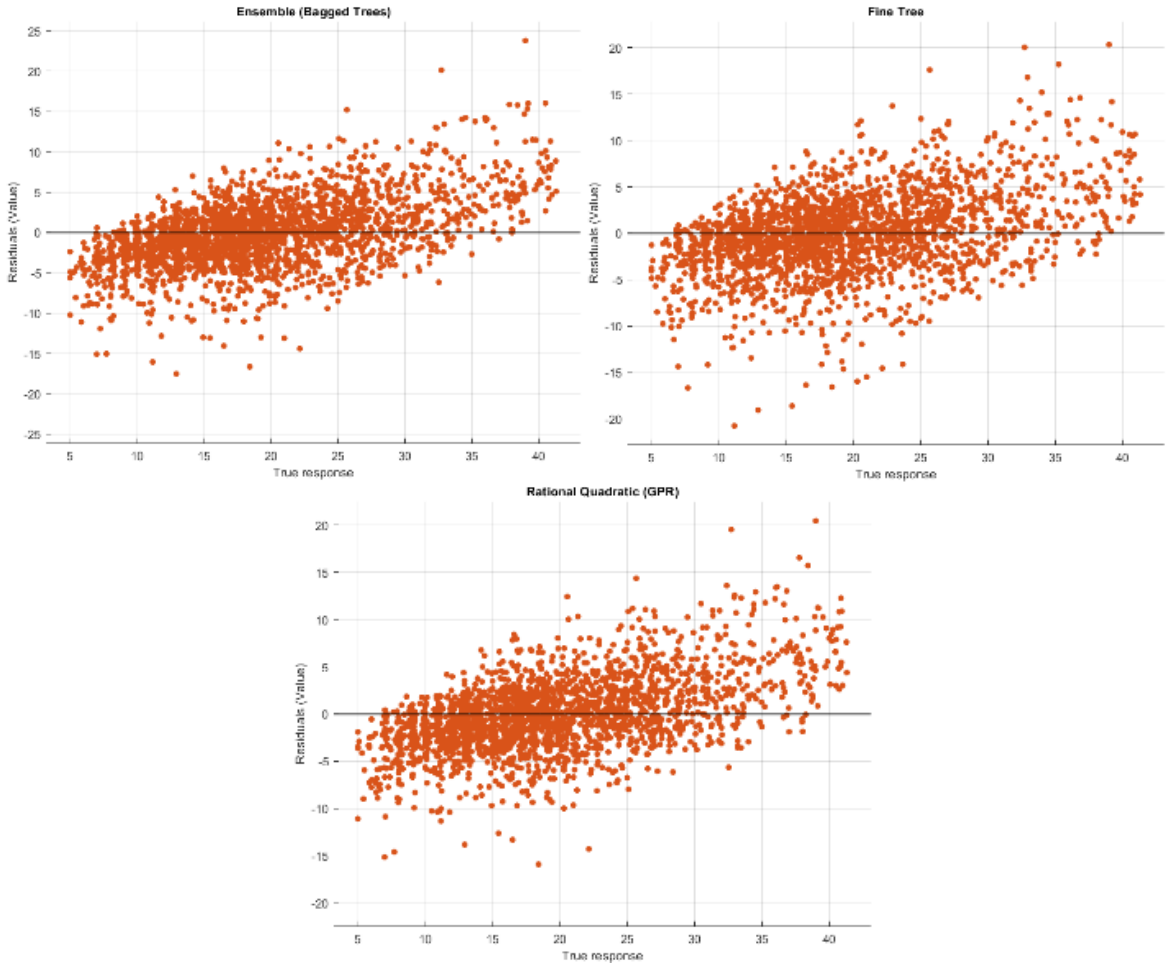

Fig. 5 Scatter plots of residuals for the testing phase

The difference between MAE and R^2 is that MAE (having the same units as the measurements), is interpreted with a knowledge of the phenomenon that we are trying to monitor. Another popular error metric for regression problems is MSE presented in (7).

Open Science Index, Environmental and Ecological Engineering Vol:18, No:7, 2024 publications.waset.org/10013700.pdf

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \quad (7)$$

where n is the number of data points, Y_i is the observed value, and $\hat{Y}_i$ is the predicted value**.**

VI. Results and Discussion

The findings indicate that the three models displayed varying levels of accuracy in their predictions. The Rational Quadratic GPR showed the results with the RMSE and MSE, on both the validation and test datasets as presented in Table II suggesting its superior ability to forecast PM10 concentrations. The Bagged Trees also delivered good performance achieving outcomes across all evaluation metrics. While the Tree model performed quite acceptably compared to the other models, it excelled in terms of speed. However. the choice of the appropriate model should consider the trade-off between computational efficiency and predictive performance based on specific application requirements. The Fine Tree model yielded an RMSE of 4.62 during the validation phase, with corresponding MSE and R^2 values of 21.34 and 0.65, respectively. The MAE stood at 3.39. In the test phase, the RMSE slightly increased to 4.33, accompanied by an MSE of 18.76 and an R^2 of 0.68. The MAE for the test set was 3.39. The response between observations and predictions is shown in Fig. 4 in the validation phase, and Fig. 5 demonstrates the residuals in the testing phase.

The Bagged Trees model showed strong performance with an RMSE of 4.40 MSE of 19.36 and R^2 of 0.68 during validation. The MAE was 3.32. In the testing phase, the model maintained its effectiveness with an RMSE of 4.20 MSE of 17.61 and an R^2 of 0.70. The MAE for the test set remained at 3.32 as well. The visual representations are shown in Figs. 4 and 5. On the hand, the GPR model exhibited strong performance during validation with an RMSE of 4.14 MSE of 17.15 and an R^2 of 0.72. The MAE was recorded at 3.08 in this phase as well.

During testing, the GPR model continued to perform with an RMSE of 3.89 and a MSE value standing at 15.13. Impressively reaching a value for R^2 equaling out to be approximately around 0.75. The MAE, for this test set also remained stable at a value to 3.08.

TABLE II
Performances of Models

Model		Fine Tree	Bagged Trees	GPR
RMSE	Validation	4.62	4.40	*4.14*
MSE		21.34	19.36	17.15
R^2		0.65	0.68	0.72
MAE		3.39	3.32	3.08
RMSE	Test	4.33	4.20	*3.89*
MSE		18.76	17.61	15.13
R^2		0.68	0.70	0.75
MAE		3.19	3.14	2.91

VII. Conclusion

In summary, this research applied three machine learning models Fine Tree, Ensemble (Bagged Trees), and GPR to predict PM10 concentrations based on hourly data collected in Vilnius, Lithuania, from January 2020 to December 2022 on around 22,867 records. Meteorological information, including wind speed, temperature, and dew point, served as predictor variables. The findings revealed varying levels of accuracy among the models. GPR demonstrated superior forecasting ability for PM10 concentrations, as evidenced by lower RMSE of 4.14 for validation and 3.89 for the testing phase. The Bagged Trees model also exhibited good performance across all evaluation metrics. The Tree model excelled in terms of speed with acceptable accuracy regarding the experimental results, emphasizing the importance of carefully selecting a model based on the specific needs of the application, taking into account the balance between computational efficiency and predictive accuracy.

VIII. Limitation

A notable disadvantage at some points is the lack of verification of AP data, which results in the exclusion of data points. Otherwise, the model could benefit from further improvement, the AP model requires proper training, validation, and testing stages. High-quality data must be carefully and precisely picked, vetted, and evaluated by comparing results to known values. Validation involves inspecting for errors, inconsistencies, and missing values. Comparing model predictions to actual air quality measurements from reliable sources to assess the model's accuracy, although the study integrates some of the key predictors of PM10 concentrations, the inclusion of additional variables such as variations in traffic flow, industrial outputs, and cross-border pollution might further enhance the models' predictive capabilities.

References

[1] Y. A. Ayturan et al., "Short-term prediction of pm2.5 pollution with deep learning methods," Global Nest Journal, vol. 22, no. 1, pp. 126–131, 2020, doi: 10.30955/gnj.003208.

[2] R. Sharma, G. Shilimkar, and S. Pisal, "Air Quality Prediction by Machine Learning," Int J Sci Res Sci Technol, pp. 486–492, May 2021, doi: 10.32628/ijsrst218396.

[3] H. K. Doreswamy, Y. Km, and I. Gad, "ScienceDirect Forecasting Air Pollution Particulate Matter (PM 2.5) Using Machine Learning Regression Models," Procedia Comput Sci, vol. 171, pp. 2057–2066, 2020, doi: 10.1016/j.procs.2020.04.221.

[4] W. C. Leong, R. O. Kelani, and Z. Ahmad, "Prediction of air pollution index (API) using support vector machine (SVM)," 2019, doi: 10.1016/j.jece.2019.103208.

[5] S. Abdullah, M. Ismail, and S. Y. Fong, "Multiple Linear Regression (MLR) models for long term Pm10 concentration forecasting during different monsoon seasons," J Sustain Sci Manag, vol. 12, pp. 60–69, Oct. 2017.

[6] K. Veljanovska and A. Dimoski, "Machine Learning Algorithms in Air Quality Index Prediction," International Journal of Science and Engineering Investigations, vol. 6, p. 71, 2017, Accessed: Oct. 15, 2023. (Online). Available: www.IJSEI.com

[7] X. Zhao, R. Zhang, J.-L. Wu, and P.-C. Chang, "A Deep Recurrent Neural Network for Air Quality Classification," Journal of Information Hiding and Multimedia Signal Processing c, vol. 9, no. 2, 2018.

[8] M. Kulkarni, A. Raut, S. Chavan, N. Rajule, and S. Pawar, "Air Quality Monitoring and Prediction using SVM," in 2022 6th International Conference on Computing, Communication, Control and Automation, ICCUBEA 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICCUBEA54992.2022.10010942.

Open Science Index, Environmental and Ecological Engineering Vol:18, No:7, 2024 publications.waset.org/10013700.pdf

[9] L. Contreras Ochando, C. I. Font Julián, and F. Contreras Ochando, "Airvlc: An application for real-time forecasting urban air pollution C` esar Ferri," 2015, Accessed: Oct. 16, 2023. (Online). Available: http://www.aemet.es/
[10] M. Aljanabi, M. Shkoukani, and M. Hijjawi, "Ground-level Ozone Prediction Using Machine Learning Techniques: A Case Study in Amman, Jordan," International Journal of Automation and Computing, vol. 17, no. 5, pp. 667–677, 2020, doi: 10.1007/s11633-020-1233-4.
[11] J. He et al., "Atmospheric Pollution Research 14 (2023) 101832 Available online 2," pp. 1309–1042, 2023, doi: 10.1016/j.apr.2023.101832.
[12] "What is Air Pollution?" Accessed: Nov. 16, 2023. (Online). Available: https://www.aqi.in/blog/what-is-air-pollution/
[13] J. Cook et al., "Quantifying the consensus on anthropogenic global warming in the scientific literature," Environmental Research Letters, vol. 8, no. 2, 2013, doi: 10.1088/1748-9326/8/2/024024.
[14] M. Kampa and E. Castanas, "Human health effects of air pollution", doi: 10.1016/j.envpol.2007.06.012.
[15] "WHO releases country estimates on air pollution exposure and health impact." Accessed: Oct. 03, 2023. (Online). Available: https://www.who.int/en/news-room/detail/27-09-2016-who-releases-country-estimates-on-air-pollution-exposure-and-health-impact
[16] R. Sher et al., "Air pollution and its health impacts in Malaysia: a review", doi: 10.1007/s11869-020-00867-x.
[17] A. Zhalehdoost and M. Taleai, "A Review of the Application of Machine Learning and Geospatial Analysis Methods in Air Pollution Prediction," Pollution, vol. 8, no. 3. University of Tehran, pp. 904–933, May 01, 2022. doi: 10.22059/POLL.2022.336044.1300.
[18] G. K. Kang, J. Z. Gao, S. Chiao, S. Lu, and G. Xie, "Air quality prediction: Big data and machine learning approaches," Int. J. Environ. Sci. Dev, vol. 9, no. 1, pp. 8–16, 2018.
[19] "Vilnius Air Quality Index." Accessed: Oct. 28, 2023. (Online). Available: https://www.iqair.com/lithuania/vilnius
[20] T. Madan, S. Sagar, and D. Virmani, "Air Quality Prediction using Machine Learning Algorithms-A Review," in Proceedings - IEEE 2020 2nd International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2020, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 140–145. doi: 10.1109/ICACCCN51052.2020.9362912.
[21] Q. Zhang, H. Yu, M. Barbiero, B. Wang, and M. Gu, "Artificial neural networks enabled by nanophotonics," Official journal of the CIOMP, pp. 2047–7538, doi: 10.1038/s41377-019-0151-0.
[22] "CS 230 - Recurrent Neural Networks Cheatsheet." Accessed: Oct. 12, 2023. (Online). Available: https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks
[23] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
[24] K. Cho et al., "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. doi: 10.3115/v1/D14-1179.
[25] R. Gove and J. Faytong, "Machine Learning and Event-Based Software Testing: Classifiers for Identifying Infeasible GUI Event Sequences," 2012, doi: 10.1016/B978-0-12-396535-6.00004-1.
[26] C. M. Bishop, Pattern Recognition and Machine Learning. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
[27] S. Buhai, "Quantile regression: overview and selected applications," Ad Astra, vol. 4, no. 4, pp. 1–17, 2005.
[28] "Air Quality e-Reporting (AQ e-Reporting)." Accessed: Mar. 10, 2024. (Online). Available: https://www.eea.europa.eu/en/datahub/datahubitem-view/3b390c9c-f321-490a-b25a-ae93b2ed80c1
[29] "National Centers for Environmental Information (NCEI)." Accessed: Mar. 10, 2024. (Online). Available: https://www.ncei.noaa.gov/
[30] Y. Feng et al., "Prediction of Hourly Air-Conditioning Energy Consumption in Office Buildings Based on Gaussian Process Regression," Energies (Basel), vol. 15, no. 13, Jul. 2022, doi: 10.3390/en15134626.
[31] C. E. Rasmussen and C. K. I. Williams, "Gaussian Processes for Machine Learning", Accessed: Mar. 12, 2024. (Online). Available: www.GaussianProcess.org/gpml
[32] T. Hastie, R. Tibshirani, and J. Friedman, "Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction."
[33] S. Guo, X. Tao, and L. Liang, "Exploring Natural and Anthropogenic Drivers of PM2.5 Concentrations Based on Random Forest Model: Beijing–Tianjin–Hebei Urban Agglomeration, China," Atmosphere (Basel), vol. 14, no. 2, Feb. 2023, doi: 10.3390/atmos14020381.
[34] "How to assess air quality sensor accuracy: MAE." Accessed: Oct. 26, 2023. (Online). Available: https://www.clarity.io/blog/how-to-assess-sensor-accuracy-mae

Open Science Index, Environmental and Ecological Engineering Vol:18, No:7, 2024 publications.waset.org/10013700.pdf