The University of Tehran Press

Pollution

Print ISSN: 2383-451X
Online ISSN: 2383-4501

https://jpoll.ut.ac.ir/

Daily PM_{10} Prediction of Thiruvananthapuram City and Interpretability Analysis of Influencing factors

Sherin Babu[1,2]✉ | Binu Thomas[2,3]

1. Department of Computer Science, Assumption College Autonomous, Changanassery, Kottayam, Kerala, India
2. School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India
3. Department of Computer Applications, Marian College, Kuttikanam, Idukki, Kerala, India

Article Info

Article type:
Research Article

Article history:
Received: 21 September 2024
Revised: 9 December 2024
Accepted: 19 January 2025

Keywords:
PM_{10}
Ensemble models
SHAP
Regression
Extra Trees

ABSTRACT

Accurate predictions of air pollutant PM10 concentrations are essential for crafting effective air quality management strategies. This study compares three decision tree ensemble models—Random Forest (RF), Extra Trees, and Extreme Gradient Boosting (XGBoost)—to forecast daily PM_{10} levels in Thiruvananthapuram, India. By integrating meteorological data and air pollutant variables, this study aims to enhance both the accuracy and interpretability of urban air pollution dynamics. Spearman correlation analysis is employed to analyse the relationships between PM_{10} and the various input features. The predictive performance of the ensemble models is evaluated using Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2). The Extra Trees model demonstrates superior predictive performance, achieving an R^2 of 0.945 and an RMSE of 8.174 μg/m³. The model-agnostic interpretability method SHapley Additive exPlanations (SHAP) demonstrates that $PM_{2.5}$, NH_3, NO_2, and O_3 have a major impact on PM_{10} forecasts. Additionally, it reveals that meteorological conditions, particularly rainfall and relative humidity, play a crucial role in determining PM_{10} concentrations. This research highlights the potential of machine learning techniques, especially when combining the Extra Trees model with SHAP, to assist local governments in strategic planning and air quality management efforts. Although temporal coverage limits are acknowledged, this study offers useful information to environmental agencies and policymakers looking for data-driven strategies to reduce air pollution.

Cite this article: Babu, Sh., & Thomas, B. (2025). Daily PM_{10} Prediction of Thiruvananthapuram City and Interpretability Analysis of Influencing factors. *Pollution*, 11(2), 525-537.
https://doi.org/10.22059/poll.2024.382674.2561

DOI: https://doi.org/10.22059/poll.2024.382674.2561

INTRODUCTION

The total number of solid and liquid particles suspended in the air, many of which are hazardous, is known as particulate matter. Inhalable particles with a diameter of 10 micrometres or less are referred to as PM_{10} (Tong et al., 2020; X. Wu et al., 2020). The formation of PM_{10} is influenced by a combination of environmental factors and anthropogenic activities. Environmental factors include meteorological conditions and natural events such as wildfires, volcanic eruptions, and dust storms (Sohrab et al., 2024). Anthropogenic activities comprise of vehicular transmissions, agricultural, industrial and construction activities (Abbas et al., 2021). Asthma, respiratory infections, lung cancer, and chronic obstructive pulmonary disease(COPD) are all caused by exposure to PM_{10} (WHO Regional Office for Europe, 2013). The elderly and children with chronic heart or lung disease are most likely to suffer negative health effects from PM_{10} exposure, according to the researchers (Brunekreef & Holgate, 2002; Pope III, 2002). Increased PM_{10} concentrations have been linked to a higher mortality rate. India has one of the highest rates

*Corresponding Author Email: *sherinbabu@assumptioncollege.edu.in*

of cardiovascular disease (CVD) prevalence worldwide. The annual number of CVD deaths in India is anticipated to rise from 2.26 million in 1990 to 4.77 million in 2025 (Huffman et al., 2011). The research study conducted about the risk factors associated with Chronic obstructive pulmonary disease (COPD), in the city of Thiruvananthapuram identified air particulate matter as a significant contributor (Surendran et al., 2022).

Thiruvananthapuram, the capital of Kerala, the southernmost state of India, is indeed distinguished by such unique geographical location along the southwestern coast, by lush green geography, and by rich cultural heritage. Also, about 1.5 million people inhabit this city, which engages itself in some of the most important roles as an education, industrial, tourist and administrative canter. The major sources of air pollution in Thiruvananthapuram are attributed to vehicular transmission, landfills, industrial emissions, waste burning and construction activities (Aiswarya et al., 2023; Kumar & Swarnalatha, 2019). When compared to a number of other major Indian cities, Thiruvananthapuram, has demonstrated a much superior air quality condition (Lavanyaa et al., 2023). Also, the city experiences fluctuations in air quality due to various factors, including seasonal changes and meteorological conditions and hence has problems with air pollution, especially with regard to PM_{10} and other contaminants (Nishanth et al., 2012; Sumesh et al., 2017).

Innovations in computational methods and the availability of large amount of data storage devices, have resulted in the development of applications for predicting air pollutant concentrations for a spectrum of uses. Machine learning algorithms have been successfully applied to the forecasting of a wide range of air pollutant concentrations over a variety of time scales (Bellinger et al., 2017; Xi et al., 2015). Researchers developed an ANN and SVM based PM_{10} forecasting model with a two-year data set of air pollutant and meteorological parameters from Taiyuan, China, and then the Taylor expansion forecasting model to revise the forecasting goal, resulting in a high accuracy rate (P. Wang et al., 2015). A random forest model that used satellite, meteorologic, atmospheric, and land-use data for predicting daily $PM_{2.5}$ concentrations at a resolution of 1 × 1 km throughout an urban area, was developed by researchers (Brokamp et al., 2018). Researchers developed Artificial Neural Networks (ANN), Boosted Regression Trees (BRT), and SVM machine learning models to predict PM_{10} and $PM_{2.5}$ levels based on traffic, meteorological, and pollutant data collected from various locations in London from 2007 to 2012(Suleiman et al., 2019). A method for integrating quantile regression into the boosted regression trees (BRT) technique for the purpose of forecasting PM_{10} in Malaysia was proposed in the research work of (Verma et al., 2024). Accurate prediction mechanism for PM_{10} and $PM_{2.5}$ was devised in Seoul, South Korea, using meteorological data and tree-based machine learning methods, and light gradient boosting method yielded the most accurate prediction results (Kim et al., 2022). The gradient-boosting regression tree model demonstrated the most effective performance in the research conducted to forecast PM_{10} concentrations in the Caribbean region (Plocoste & Laventure, 2023).

While many studies have used various machine learning models to predict PM_{10}, these black-box models often fail to identify the factors affecting forecasting accuracy. Understanding these variables is crucial for improving system efficacy and minimizing costs in air pollution prediction. Based on the literature, it is noted that the researchers employed explainable AI frameworks such as Permutation Feature Importance (PFI) and SHapley Additive exPlanations (SHAP) to interpret the output of machine learning models, thereby addressing the challenges posed by black-box models. XGBoost and SHAP were employed to investigate the influence of meteorological factors on PM_{10} concentrations in the Belgrade region of Serbia (Stojić, 2021). A random forest model with SHAP was implemented to investigate the spatiotemporal fluctuations of meteorological, socioeconomic, topographic, and land cover factors that are affecting the concentrations of $PM_{2.5}$ in Zhejiang Province, China (Li et al., 2021). The research work of (Y. Wu et al., 2022) described the seasonal prediction of $PM_{2.5}$ concentrations in Beijing using a variety of machine learning models, as well as the impact of meteorological factors on the specific predictions, using SHAP. In the research investigation of (S. Wang et al., 2023), a machine learning interpretation method based on SHAP was proposed to analyse the factors that contribute to the variation of $PM_{2.5}$ and O_3 concentrations, based on the CatBoost model. The evaluation of human and meteorological influences on PM_{10} predictions for Queensland, Australia was conducted using a variety of decision tree ensemble models and SHAP (Verma et al., 2024).

Despite significant advancements in understanding air quality dynamics and the application of conventional statistical techniques for pollutant prediction, there remains a notable gap in utilizing the effectiveness of ML methodologies for predicting PM_{10} concentrations in the study region. In recent years, there has been a growing interest in understanding PM_{10} due to its significant health impacts and environmental implications. Most existing studies in the study region have primarily focused on characterizing PM_{10}, its health impacts, identifying its various physio-chemical properties, and determining the sources of its origination. Even while some progress has been made in this regard, there is still a large gap in the development and implementation of machine learning models designed especially to forecast PM_{10} concentrations in the study area. Moreover, although tools such as SHAP have been widely used to improve model interpretability, limited research work systematically explores how these analysis results can guide decision processes in air quality management.

Hence this research work aims to explore and define key aspects related to PM_{10} prediction within the framework of decision tree ensemble methods, emphasizing their interdependencies with meteorological and air pollutant factors. This is the initial study to employ machine learning and ML based interpretability methods to elucidate the factors that affect the PM_{10} concentrations of Thiruvananthapuram city. The objective of this study is to examine the potential of three decision tree ensemble regression models (RF, Extra Trees and XGBoost in predicting daily PM_{10} concentrations in Thiruvananthapuram city, using a variety of air pollutant and meteorological factors as input features. Using SHAP analysis, the study then seeks to identify the influential factors of PM_{10} prediction from the well-performing decision tree ensemble model. The research objectives of this study are

(1) To analyse the performance of decision tree ensemble models in PM_{10} prediction.
(2) To identify the key features influencing PM_{10} concentrations using SHAP analysis.
(3) To enhance the understanding of feature contributions to PM_{10} predictions.

MATERIALS AND METHODS

Study Area and Dataset

This study is conducted for the capital city of Kerala, Thiruvananthapuram. The ambient air quality monitoring station in the city is located at Plammoodu (Latitude: 8.51N, Longitude: 76.94). Kerala State Pollution Control Board (KSPCB) owns and operates the monitoring station. The daily data for the analysis is obtained from the Central Pollution Control Board’s (CPCB) website. The data is collected for 914 days, from July 1, 2017 to December 31, 2019. The dataset contains daily values of PM_{10} and other air contaminants termed as $PM_{2.5}$, NO, NO_2, NO_x, NH_3, CO, O_3, and SO_2. Wind speed (WS), wind direction (WD), atmosphere air temperature (AT), relative humidity (RH), rainfall volume (RF), ambient temperature (Temp), solar radiance (SR), and buoyancy pressure (BP) are the meteorological parameters included in the data set. PM_{10} is the target variable and the remaining 16 variables are the independent variables. The machine learning models and model interpretability technique SHAP in this study are developed using Colab Notebook, a Google Cloud Computing service, in Python programming language.

In the data preprocessing phase, records containing missing values are excluded from the dataset, as the quantity of missing values is minimal (Blenkinsop et al., 2015; Kujawska et al., 2022). The kernel density estimate (KDE) plot of PM_{10}, which is a graphical representation that estimates the probability density function of target variable PM_{10} is shown in Figure 1. The plot exhibits a single, prominent peak around a PM_{10} value of around 50-60 µg/m³. This suggests that the majority of the PM_{10} measurements fall within this range. Also, the distribution appears to be relatively symmetric, with the peak located close to the centre of the X-axis. This indicates a relatively normal or Gaussian-like distribution of PM_{10} concentrations. The tails of the distribution extend towards both lower and higher PM_{10} values, suggesting the presence of some outliers or less common observations at the extremes. The width of the distribution provides an indication of the overall variability or dispersion of PM_{10} levels in the dataset.

Here the Spearman correlation is used to analyse the association between the target variable PM_{10} and the input features. It is a useful tool for assessing the strength and direction of monotonic relationships between variables (Alsaqr, 2021). Spearman correlation heatmap of the features used in this study is shown in Figure 2. Every square shows the correlation outcome of two different variables. The heatmap establishes that $PM_{2.5}$ has the highest positive correlation to PM_{10}. The air pollutants O_3, NH_3, SO_2 and NO_2 also have significant positive correlation with the PM_{10}. The meteorological factors that show moderate positive correlation with PM_{10} are SR, BP and ambient temperature. The RH and RF features exhibit a substantial negative correlation with the PM_{10}, whereas the wind speed exhibits a mild negative correlation.

The Spearman correlation heatmap demonstrates that the majority of air pollutant factors have significant direct effects on PM_{10}, while meteorological conditions are not having much significant impact. The correlation heatmap emphasizes the presence of multicollinearity, particularly among air pollutant factors. The impact of explanatory variables on the target variable is difficult to comprehend when relying solely on correlation coefficient matrices due to multicollinearity, which complicates explanatory analysis using traditional methods. SHAP method can be implemented to resolve this issue. This method improves interpretability by elucidating the influence mechanisms, even in the presence of multicollinearity.

Random Forest (RF) Regressor

RF is a type of supervised machine learning model that integrates several decision trees

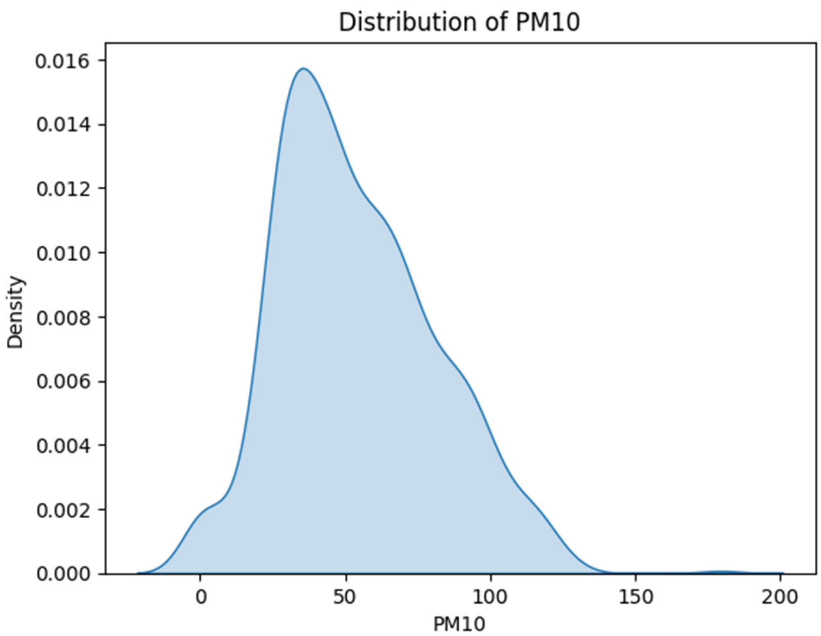

Fig. 1. Visualization of the PM_{10} data distribution

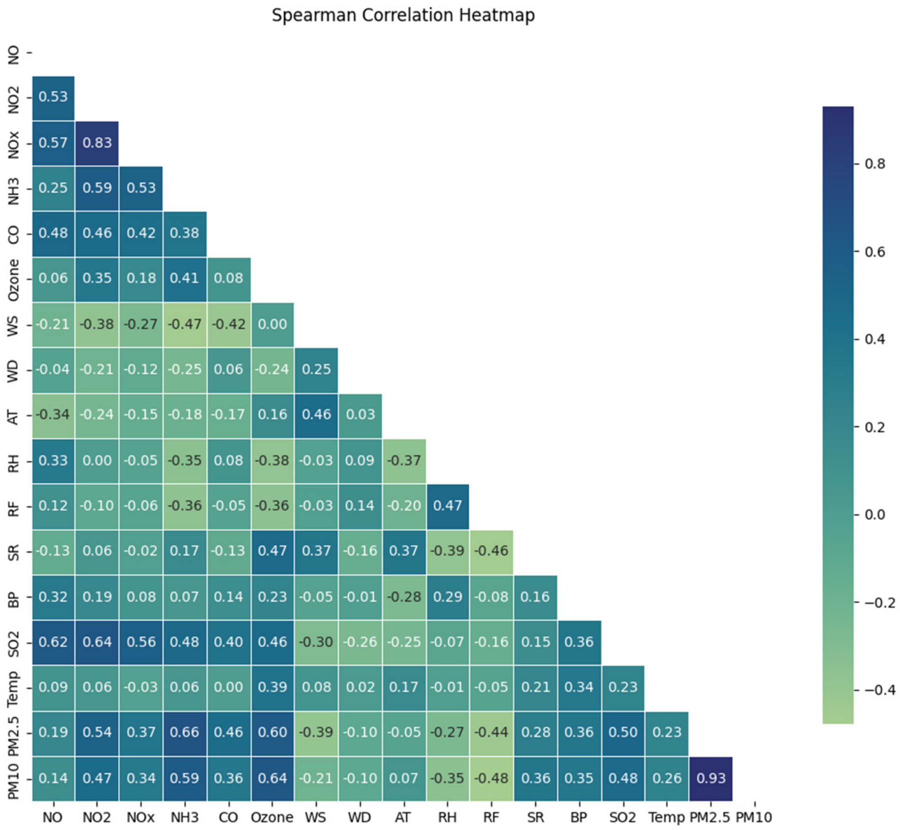

Fig. 2. Spearman Correlation heatmap of all variables

into a single model for numerical value prediction. It is an ensemble model that tries to lessen overfitting and at the same time attempts to augment accuracy by combining the predictions of several decision trees (Kabiraj et al., 2020). Each tree in the forest uses a different subset of the data as the basis for its own independent forecast. The final input prediction is based on the average, of all the predictions provided by each individual tree. RF works on the concept of integrating numerous decision trees to determine the final output instead of depending solely on individual decision trees. RF method is based on Bootstrap and Aggregation, often known as bagging(Prasad et al., 2021; J. Zhou et al., 2020).

Extra Trees Regressor

Extra Trees regressor, also known as Extremely Randomised Trees, is a kind of ensemble learning technique that generates a forecast by combining the output of several de-correlated decision trees gathered in a “forest.”(Geurts et al., 2006). Unlike RF, decision trees are trained using the complete dataset in Extra Trees. It is substantially faster than RF, because Extra Trees uses a random algorithm to choose the value at which to split features rather than RF’s greedy technique (Wehenkel et al., 2006; Yarveicy & Ghiasi, 2017). Extra Trees includes creating a randomised ensemble of trees and aggregating their predictions in a suitable manner, such as arithmetic or majority voting in classification/regression problems (Nistane & Harsha, 2018; Seyyedattar et al., 2020). It applies the random forest principle, by training each base estimator with a random subset of features. However, when splitting the node, it chooses the best function and the corresponding value at random. The cut points used to break nodes in Extra Trees and RF are also different. Extra Trees selects the best split when Random Forest chooses it at random.

Extreme Gradient Boosting (XGBoost) Regressor

XGBoost is a supervised machine learning algorithm which is used to make predictions on continuous numerical data. It makes use of the gradient boosting ensemble method, which builds a stronger, more accurate model by combining the predictions of several weaker models (Asselman et al., 2023). An ensemble of decision trees is produced by XGBoost, and each tree is trained to generate predictions using a subset of the given data. The trees are grown one after the other, each one picking up insights from its predecessor's errors. The average of the forecasts from each tree in the ensemble is used to get the final prediction (Lin et al., 2022; L. Zhang et al., 2020). XGBoost's efficiency in managing big datasets and missing data is one of its advantages.

Performance Evaluation

The root mean squared error (RMSE) and coefficient of determination (R^2) are employed to evaluate the performance of the established tree models in predicting PM_{10}. The association between the actual and predicted PM_{10} values is calculated using the determination coefficient, R^2. An R^2 of 1 indicates that the model predictions perfectly fit the data. The RMSE is a metric that quantifies the average variation between the predicted and actual values of a model. It offers an estimate of the model's ability to accurately predict the objective value. The model is more accurate when the RMSE is lower.

Model Interpretability using SHAP

SHapley Additive exPlanations (SHAP) is a frequently employed approach for interpreting predictions in black box type machine learning models and is developed by Lund berg and Lee (Lundberg & Lee, 2017). This is a model-agnostic technique and can be applied to a wide variety of machine learning models (Chaibi et al., 2021; Ullah et al., 2023). Over the past few years, there has been an increasing interest in the application of SHAP to elucidate machine learning models. The SHAP method is founded on Shapley values in cooperative game theory, which are used to evaluate the contributions of each participant in a game (Li et al., 2021; Rajput et al., 2023). The objective is to distribute benefits equitably among players who join a coalition. The relationship between Shapley values and model interpretation is derived from the fact that the variables used for training are referred to as "players," while the model's predictions represent the matching "revenues" (Y. Zhang et al., 2024). The SHAP method enables users to gain a deeper understanding of the importance of individual variables in predicting outcomes, thereby facilitating a more comprehensive understanding of sophisticated machine learning models. SHAP offers both local and global explanations for machine learning models (Zheng et al., 2023). The SHAP Python module and the TreeExplainer library are employed to generate SHAP interpretations in this study.

RESULTS AND DISCUSSION

The potential of the three ensemble models in PM_{10} prediction is examined. Here the tree models are fitted on the training dataset with 80% of the original dataset and then tested on the test dataset with the remaining 20% of the original dataset (Bharat et al., 2018; Bhatt et al., 2023). The default parameter setting of RF, Extra Trees and XGBoost models are employed. The results of performance metrics of the three models are given in Table 1.

Higher R^2 values indicate better model performance in explaining variability in PM_{10} concentrations and all models show strong explanatory power, with Extra Trees model leading (Puri et al., 2018). Lower RMSE values indicate more accurate predictions (Ağbulut et al., 2021). It is evident that Extra Trees model resulted in the least RMSE (8.174 $\mu g/m^3$) and the highest R^2 score (0.945), compared to the RF model and XGBoost models. This indicates that

Table 1. Performance metrics of PM_{10} prediction

Model	R^2	*RMSE ($\mu g/m^3$)*
Extra Trees	0.945	8.174
Random Forest	0.939	8.655
XGBoost	0.929	8.833

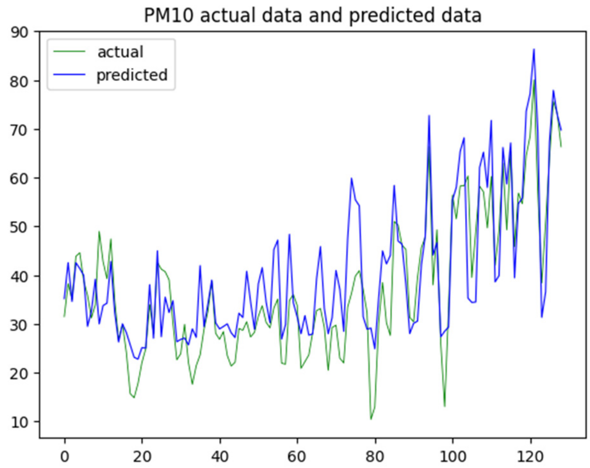

Fig. 3. Plot of actual and predicted PM_{10} values by Extra Trees model

approximately 94.5% of the variance in PM_{10} concentrations can be explained by the Extra Trees model. An RMSE of 8.174 indicates that, on average, the Extra Trees model predictions deviate from actual PM_{10} values by about 8.17 µg/m³. Hence Extra Trees model outperforms both Random Forest and XGBoost in terms of R^2 and RMSE, indicating it provides the best balance between accuracy and precision for PM_{10} prediction among the models evaluated.

The plot of PM_{10} predicted values vs. actual PM_{10} values using the best performing Extra Trees model is shown in Figure 3. The blue line represents the predicted PM_{10} measurements, while the green line shows the actual PM_{10} values. The predicted PM_{10} values generally follow the overall trend of the actual PM_{10} measurements over the test data and it indicates that the model used for the predictions is able to capture the broad patterns and dynamics of the PM_{10} concentrations. The discrepancies in the plot represent situations where the Extra Trees model's predictions do not align perfectly with the actual PM_{10} concentrations.

The prediction residual plot based on Extra Trees regression model is shown in Figure 4. Majority of the data residuals (the difference between the actual and predicted values) shown in the figure are close to the zero baseline, which proves that the developed Extra Trees model provides a good prediction of PM_{10} values. Also, the residuals are scattered around the horizontal zero line, indicating that the model's predictions are generally unbiased (Espinheira et al., 2021). The magnitude of the residuals is generally within the range of -20 to +20 µg/m³, suggesting that the model's predictions have a reasonably good fit to the actual PM_{10} values.

In order to evaluate the influence of predictor variables on the predictions made by the machine learning algorithms, SHAP technique is implemented. SHAP is established on the Extra Trees prediction model, which demonstrates the highest level of prediction performance. Valuable insights into the impact of factors on forecasting outcomes are provided by the SHAP results, which captures both individual factor effects and relationships among factors. The beeswarm plot is employed to demonstrate the global contribution of each individual feature on the model's predictions, as shown in Figure 5.

In beeswarm plot, each row represents an individual feature and the features are organized

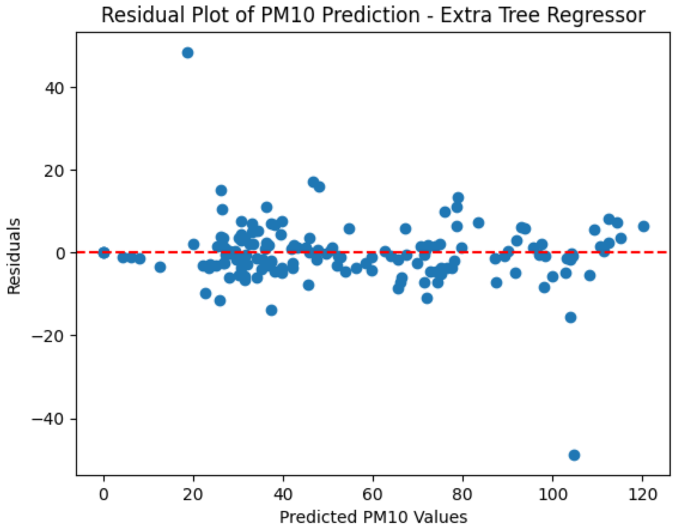

Fig. 4. Residual Plot of PM_{10} prediction by Extra Trees model

in a hierarchy with the decreasing order of importance. The beeswarm plot provides a concise overview of the magnitude and direction of each attribute's global impact. The horizontal position of the dots indicates whether the feature has a positive or negative impact on PM_{10} prediction. The dots to the right side indicate a positive impact, while those to the left side indicate a negative one. The colours of the dot denote the value of the feature, which assists in determining its impact on PM_{10} predictions. The feature's higher values are represented by red points, while the lower values are represented by blue points. The density of the dots indicates the degree to which each feature influences PM_{10} predictions across a variety of data points. It is shown that $PM_{2.5}$ is the greatest contributor to the PM_{10} formation. Higher values of $PM_{2.5}$ contribute to higher PM_{10} concentrations (Dongarrà et al., 2010). Following $PM_{2.5}$, NH_3, NO_2, and O_3 also demonstrate a higher level of positive significance in the forecasted PM_{10} results (Huang et al., 2021; Riches et al., 2022). Air pollutants NO_x and SO_2 are having acceptable contribution in predicting PM_{10}. However, among all the air pollutant factors, NO and CO have the least significant effect. In terms of meteorological conditions, the most significant impact on PM_{10} is done by the factors RF and RH, with BP and SR following as the next most influential factors. Higher values of RF, wind speed and wind direction result in lesser PM_{10} values. Rainfall largely lowers PM_{10} concentrations through the washout effect, in which raindrops absorb and remove suspended particles from the environment (Y. Zhou et al., 2020). High humidity can raise PM_{10} concentrations by increasing their likelihood of remaining suspended in the air (Li et al., 2017). However, ambient temperature and air temperature indicate only a negligible impact on PM_{10}.

The SHAP technique can also elucidate the local interpretation of influence of features on PM_{10} prediction, quantify the relative importance of features, and explain the outcomes of individual observations. Figure 6 illustrates the SHAP method's explanation on a single instance of PM_{10} prediction (with PM_{10} value = 40.30), that is arbitrarily selected from the test dataset. It can be seen that $PM_{2.5}$ demonstrated the highest contribution, followed by NH_3 and O_3. The local interpretations are aggregated by averaging the absolute Shapley values per attribute across the data in order to generate a global interpretation of the Extra Trees model predictions. The SHAP feature importance plot, depicted in Figure 7, illustrates the global impact of each feature on the prediction of PM_{10}. The SHAP method reveals that the most critical features are $PM_{2.5}$ and NH_3, while the least effective feature is air temperature. NO_2, O_3, NO_x, and SO_2 are among the most significant air pollutants that influence PM_{10} predictions, while RF and RH are

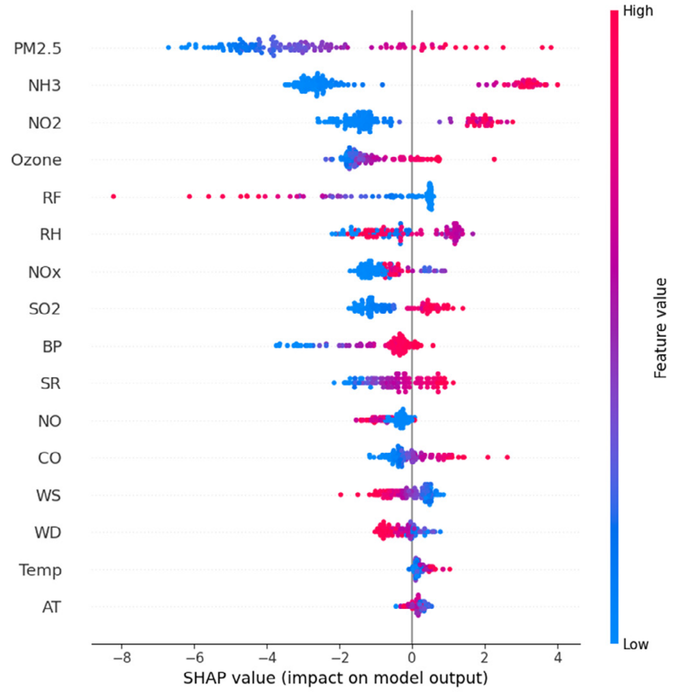

Fig. 5. Beeswarm summary plot of SHAP

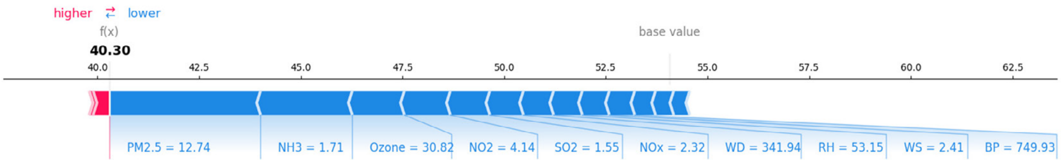

Fig. 6. Explanation of the Extra Trees model's PM_{10} output value of 40.30 using SHAP

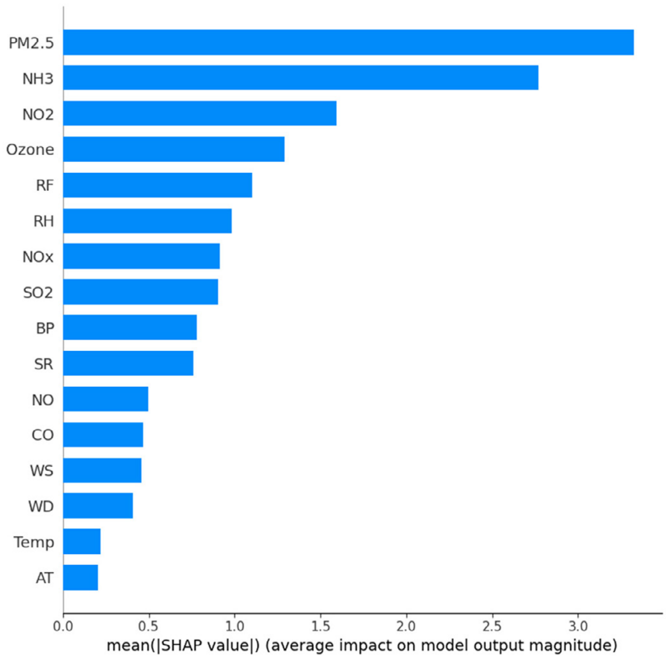

Fig. 7. SHAP Feature importance plot

the most significant meteorological factors. This is in accordance with the scientific knowledge and results obtained in the Spearman correlation analysis, which suggest a direct correlation between the aforementioned factors and PM_{10}.

CONCLUSION

PM_{10} is a significant contributor to air pollution, posing risks to both human health and the environment. Regular estimation of PM_{10} levels is essential for assessing air quality and implementing effective mitigation strategies. This study highlights the role of machine learning techniques, especially the Extra Trees model, in improving the prediction of PM_{10} concentrations in Thiruvananthapuram. By effectively combining air pollutant and meteorological data, this model outperforms other ensemble methods - Random Forest and XGBoost. The use of SHAP for interpretability analysis indicates that $PM_{2.5}$, NH_3, and NO_2 are key contributors to PM_{10} levels, underscoring the necessity for targeted regulatory actions to reduce these pollutants. Additionally, the identification of relative humidity and rainfall as influential meteorological factors highlights the importance of incorporating weather data into air quality models for improved prediction accuracy. One significant limitation of this study is its reliance on a 2.5-year time frame for predicting PM_{10} levels, which may not fully reflect the long-term environmental changes and variations in air quality. Additionally, the research does not account for fluctuations in pollutant emissions that can occur due to factors such as festivals, industrial activities, or seasonal changes. Future studies are aimed to extend both the time frame and incorporate seasonal elements to achieve a more thorough understanding of air quality dynamics. Overall, this research provides important insights for policymakers and environmental agencies, supporting informed decision-making to enhance public health and effectively address air pollution issues.

GRANT SUPPORT DETAILS

The present research did not receive any financial support.

CONFLICT OF INTEREST

The authors declare that there is not any conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/ or falsification, double publication and/or submission, and redundancy has been completely observed by the authors.

LIFE SCIENCE REPORTING

No life science threat was practiced in this research.

REFERENCES

Abbas, R., Shehata, N., Mohamed, E. A., Salah, H., & Abdelzaher, M. (2021). Environmental safe disposal of cement kiln dust for the production of geopolymers. *Egyptian Journal of Chemistry*. https://doi.org/10.21608/ejchem.2021.89060.4276

Ağbulut, Ü., Gürel, A. E., & Biçen, Y. (2021). Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. *Renewable and Sustainable Energy Reviews*, *135*, 110114. https://doi.org/10.1016/j.rser.2020.110114

Aiswarya, R., Resmi, A. R., Rahsha, C. T., Dharman, S., Adarsh, S., & Mamatha, M. (2023). Analyzing the Effect of Air Pollutants on Particulate Matter Concentrations of the Tropical coastal city of

Thiruvananthapuram, India by Wavelet Coherence. *IOP Conference Series: Earth and Environmental Science*, *1237*(1), 012017. https://doi.org/10.1088/1755-1315/1237/1/012017

Alsaqr, A. M. (2021). Remarks on the use of Pearson's and Spearman's correlation coefficients in assessing relationships in ophthalmic data. *African Vision and Eye Health*, *80*(1), Article 1. https://doi.org/10.4102/aveh.v80i1.612

Asselman, A., Khaldi, M., & Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, *31*(6), 3360–3379. https://doi.org/10.1080/10494820.2021.1928235

Bellinger, C., Mohomed Jabbar, M. S., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*, *17*(1), 907. https://doi.org/10.1186/s12889-017-4914-3

Bharat, A., Pooja, N., & Reddy, R. A. (2018). Using Machine Learning algorithms for breast cancer risk prediction and diagnosis. *2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C)*, 1–4. https://doi.org/10.1109/CIMCA.2018.8739696

Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms*, *16*(2), Article 2. https://doi.org/10.3390/a16020088

Blenkinsop, S., Chan, S. C., Kendon, E. J., Roberts, N. M., & Fowler, H. J. (2015). Temperature influences on intense UK hourly precipitation and dependency on large-scale circulation. *Environmental Research Letters*, *10*(5), 054021. https://doi.org/10.1088/1748-9326/10/5/054021

Brokamp, C., Jandarov, R., Hossain, M., & Ryan, P. (2018). Predicting Daily Urban Fine Particulate Matter Concentrations Using a Random Forest Model. *Environmental Science & Technology*, *52*(7), 4173–4179. https://doi.org/10.1021/acs.est.7b05381

Brunekreef, B., & Holgate, S. T. (2002). Air pollution and health. *The Lancet*, *360*(9341), 1233–1242. https://doi.org/10.1016/S0140-6736(02)11274-8

Chaibi, M., Benghoulam, E. M., Tarik, L., Berrada, M., & Hmaidi, A. E. (2021). An Interpretable Machine Learning Model for Daily Global Solar Radiation Prediction. *Energies*, *14*(21), Article 21. https://doi.org/10.3390/en14217367

Dongarrà, G., Manno, E., Varrica, D., Lombardo, M., & Vultaggio, M. (2010). Study on ambient concentrations of PM10, PM10–2.5, PM2.5 and gaseous pollutants. Trace elements and chemical speciation of atmospheric particulates. *Atmospheric Environment*, *44*(39), 5244–5257. https://doi.org/10.1016/j.atmosenv.2010.08.041

Espinheira, P. L., Silva, L. C. M., & Cribari-Neto, F. (2021). Bias and variance residuals for machine learning nonlinear simplex regressions. *Expert Systems with Applications*, *185*, 115656. https://doi.org/10.1016/j.eswa.2021.115656

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Huang, X., Zhang, J., Zhang, W., Tang, G., & Wang, Y. (2021). Atmospheric ammonia and its effect on PM2.5 pollution in urban Chengdu, Sichuan Basin, China. *Environmental Pollution*, *291*, 118195. https://doi.org/10.1016/j.envpol.2021.118195

Huffman, M. D., Prabhakaran, D., Osmond, C., Fall, C. H. D., Tandon, N., Lakshmy, R., Ramji, S., Khalil, A., Gera, T., Prabhakaran, P., Biswas, S. K. D., Reddy, K. S., Bhargava, S. K., & Sachdev, H. S. (2011). Incidence of Cardiovascular Risk Factors in an Indian Urban Cohort. *Journal of the American College of Cardiology*, *57*(17), 1765–1774. https://doi.org/10.1016/j.jacc.2010.09.083

Kabiraj, S., Raihan, M., Alvi, N., Afrin, M., Akter, L., Sohagi, S. A., & Podder, E. (2020). Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–4. https://doi.org/10.1109/ICCCNT49239.2020.9225451

Kim, B.-Y., Lim, Y.-K., & Cha, J. W. (2022). Short-term prediction of particulate matter (PM10 and PM2.5) in Seoul, South Korea using tree-based machine learning algorithms. *Atmospheric Pollution Research*, *13*(10), 101547. https://doi.org/10.1016/j.apr.2022.101547

Kujawska, J., Kulisz, M., Oleszczuk, P., & Cel, W. (2022). Machine Learning Methods to Forecast the Concentration of PM10 in Lublin, Poland. *Energies*, *15*(17), Article 17. https://doi.org/10.3390/en15176428

Kumar, R. S., & Swarnalatha, K. (2019). Prediction of vehicular exhaust emission for Thiruvananthapuram city. In *Recent Advances in Materials, Mechanics and Management*. CRC Press.

Lavanyaa, V. P., Harshitha, K. M., Beig, G., & Srikanth, R. (2023). Background and baseline levels of

PM2.5 and PM10 pollution in major cities of peninsular India. *Urban Climate*, *48*, 101407. https://doi.org/10.1016/j.uclim.2023.101407

Li, X., Ma, Y., Wang, Y., Liu, N., & Hong, Y. (2017). Temporal and spatial analyses of particulate matter (PM10 and PM2.5) and its relationship with meteorological parameters over an urban city in northeast China. *Atmospheric Research*, *198*, 185–193. https://doi.org/10.1016/j.atmosres.2017.08.023

Li, X., Wu, C., Meadows, M. E., Zhang, Z., Lin, X., Zhang, Z., Chi, Y., Feng, M., Li, E., & Hu, Y. (2021). Factors Underlying Spatiotemporal Variations in Atmospheric PM2.5 Concentrations in Zhejiang Province, China. *Remote Sensing*, *13*(15), Article 15. https://doi.org/10.3390/rs13153011

Lin, L., Liang, Y., Liu, L., Zhang, Y., Xie, D., Yin, F., & Ashraf, T. (2022). Estimating PM2.5 Concentrations Using the Machine Learning RF-XGBoost Model in Guanzhong Urban Agglomeration, China. *Remote Sensing*, *14*(20), Article 20. https://doi.org/10.3390/rs14205239

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Murray, C. J., & Lopez, A. D. (1997). Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study. *The Lancet*, *349*(9064), 1498–1504. https://doi.org/10.1016/S0140-6736(96)07492-2

Nishanth, T., Praseed, K. M., Rathnakaran, K., Satheesh Kumar, M. K., Ravi Krishna, R., & Valsaraj, K. T. (2012). Atmospheric pollution in a semi-urban, coastal region in India following festival seasons. *Atmospheric Environment*, *47*, 295–306. https://doi.org/10.1016/j.atmosenv.2011.10.062

Nistane, V., & Harsha, S. (2018). Performance evaluation of bearing degradation based on stationary wavelet decomposition and extra trees regression. *World Journal of Engineering*, *15*(5), 646–658. https://doi.org/10.1108/WJE-12-2017-0403

Plocoste, T., & Laventure, S. (2023). Forecasting PM10 Concentrations in the Caribbean Area Using Machine Learning Models. *Atmosphere*, *14*(1), Article 1. https://doi.org/10.3390/atmos14010134

Pope III, C. A. (2002). Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. *JAMA*, *287*(9), 1132. https://doi.org/10.1001/jama.287.9.1132

Prasad, N. R., Patel, N. R., & Danodia, A. (2021). Crop yield prediction in cotton for regional level using random forest approach. *Spatial Information Research*, *29*(2), 195–206. https://doi.org/10.1007/s41324-020-00346-6

Puri, N., Prasad, H. D., & Jain, A. (2018). Prediction of Geotechnical Parameters Using Machine Learning Techniques. *Procedia Computer Science*, *125*, 509–517. https://doi.org/10.1016/j.procs.2017.12.066

Rajput, S., Kapdi, R. A., Raval, M. S., & Roy, M. (2023). Interpretable machine learning model to predict survival days of malignant brain tumor patients. *Machine Learning: Science and Technology*, *4*(2), 025025. https://doi.org/10.1088/2632-2153/acd5a9

Riches, N. O., Gouripeddi, R., Payan-Medina, A., & Facelli, J. C. (2022). K-means cluster analysis of cooperative effects of CO, NO2, O3, PM2.5, PM10, and SO2 on incidence of type 2 diabetes mellitus in the US. *Environmental Research*, *212*, 113259. https://doi.org/10.1016/j.envres.2022.113259

Seyyedattar, M., Ghiasi, M. M., Zendehboudi, S., & Butt, S. (2020). Determination of bubble point pressure and oil formation volume factor: Extra trees compared with LSSVM-CSA hybrid and ANFIS models. *Fuel*, *269*, 116834. https://doi.org/10.1016/j.fuel.2019.116834

Sohrab, S., Csikós, N., & Szilassi, P. (2024). Effect of geographical parameters on PM10 pollution in European landscapes: A machine learning algorithm-based analysis. *Environmental Sciences Europe*, *36*(1), 152. https://doi.org/10.1186/s12302-024-00972-z

Stojić, A. (2021). Meteorological Factors Governing Particulate Matter Distribution in an Urban Environment. *Sinteza 2021 - International Scientific Conference on Information Technology and Data Related Research*, 89–93. https://doi.org/10.15308/Sinteza-2021-89-93

Suleiman, A., Tight, M. R., & Quinn, A. D. (2019). Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM10 and PM2.5). *Atmospheric Pollution Research*, *10*(1), 134–144. https://doi.org/10.1016/j.apr.2018.07.001

Sumesh, R. K., Rajeevan, K., Resmi, E. A., & Unnikrishnan, C. K. (2017). Particulate Matter Concentrations in the Southern Tip of India: Temporal Variation, Meteorological Influences, and Source Identification. *Earth Systems and Environment*, *1*(2), 13. https://doi.org/10.1007/s41748-017-0015-9

Surendran, S., Mohan, A., Valamparampil, M. J., Nair, S., Balakrishnan, S. K., Laila, A. A., Reghunath, R., Jose, C., Rajeevan, A., Vasudevakaimal, P., Surendrannair, A. T., Nujum, Z. T., Varghese, S., &

Mohan, A. (2022). Spatial analysis of chronic obstructive pulmonary disease and its risk factors in an urban area of Trivandrum, Kerala, India. *Lung India*, *39*(2), 110. https://doi.org/10.4103/lungindia.lungindia_454_21

Tong, X., Ho, J. M. W., Li, Z., Lui, K.-H., Kwok, T. C., Tsoi, K. K., & Ho, K. F. (2020). Prediction model for air particulate matter levels in the households of elderly individuals in Hong Kong. *Science of The Total Environment*, *717*, 135323.

Ullah, I., Liu, K., Yamamoto, T., Zahid, M., & Jamal, A. (2023). Modeling of machine learning with SHAP approach for electric vehicle charging station choice behavior prediction. *Travel Behaviour and Society*, *31*, 78–92. https://doi.org/10.1016/j.tbs.2022.11.006

Verma, P., Verma, R., Mallet, M., Sisodiya, S., Zare, A., Dwivedi, G., & Ristovski, Z. (2024). Assessment of human and meteorological influences on PM10 concentrations: Insights from machine learning algorithms. *Atmospheric Pollution Research*, *15*(6), 102123. https://doi.org/10.1016/j.apr.2024.102123

Wang, P., Liu, Y., Qin, Z., & Zhang, G. (2015). A novel hybrid forecasting model for PM10 and SO2 daily concentrations. *Science of The Total Environment*, *505*, 1202–1212. https://doi.org/10.1016/j.scitotenv.2014.10.078

Wang, S., Ren, Y., & Xia, B. (2023). PM2.5 and O3 concentration estimation based on interpretable machine learning. *Atmospheric Pollution Research*, *14*(9), 101866. https://doi.org/10.1016/j.apr.2023.101866

Wehenkel, L., Ernst, D., & Geurts, P. (2006). *Ensembles of extremely randomized trees and some generic applications*. https://orbi.uliege.be/handle/2268/13447

WHO Regional Office for Europe. (2013). *Review of evidence on health aspects of air pollution – REVIHAAP Project: Technical Report*. WHO Regional Office for Europe. http://www.ncbi.nlm.nih.gov/books/NBK361805/

Wu, X., Wang, Y., He, S., & Wu, Z. (2020). PM 2.5/PM 10 ratio prediction based on a long short-term memory neural network in Wuhan, China. *Geoscientific Model Development*, *13*(3), 1499–1511.

Wu, Y., Lin, S., Shi, K., Ye, Z., & Fang, Y. (2022). Seasonal prediction of daily PM2.5 concentrations with interpretable machine learning: A case study of Beijing, China. *Environmental Science and Pollution Research*, *29*(30), 45821–45836. https://doi.org/10.1007/s11356-022-18913-9

Xi, X., Wei, Z., Xiaoguang, R., Yijie, W., Xinxin, B., Wenjun, Y., & Jin, D. (2015). A comprehensive evaluation of air pollution prediction improvement by a machine learning method. *2015 IEEE International Conference on Service Operations And Logistics, And Informatics (SOLI)*, 176–181. https://doi.org/10.1109/SOLI.2015.7367615

Yarveicy, H., & Ghiasi, M. M. (2017). Modeling of gas hydrate phase equilibria: Extremely randomized trees and LSSVM approaches. *Journal of Molecular Liquids*, *243*, 533–541. https://doi.org/10.1016/j.molliq.2017.08.053

Zhang, L., Ji, Y., Liu, T., & Li, J. (2020). PM2.5 Prediction Based on XGBoost. *2020 7th International Conference on Information Science and Control Engineering (ICISCE)*, 1011–1014. https://doi.org/10.1109/ICISCE50968.2020.00207

Zhang, Y., Sun, Q., Liu, J., & Petrosian, O. (2024). Long-Term Forecasting of Air Pollution Particulate Matter (PM2.5) and Analysis of Influencing Factors. *Sustainability*, *16*(1), Article 1. https://doi.org/10.3390/su16010019

Zheng, G., Zhang, Y., Yue, X., & Li, K. (2023). Interpretable prediction of thermal sensation for elderly people based on data sampling, machine learning and SHapley Additive exPlanations (SHAP). *Building and Environment*, *242*, 110602. https://doi.org/10.1016/j.buildenv.2023.110602

Zhou, J., Asteris, P. G., Armaghani, D. J., & Pham, B. T. (2020). Prediction of ground vibration induced by blasting operations through the use of the Bayesian Network and random forest models. *Soil Dynamics and Earthquake Engineering*, *139*, 106390. https://doi.org/10.1016/j.soildyn.2020.106390

Zhou, Y., Yue, Y., Bai, Y., & Zhang, L. (2020). Effects of Rainfall on PM2.5 and PM10 in the Middle Reaches of the Yangtze River. *Advances in Meteorology*, *2020*(1), 2398146. https://doi.org/10.1155/2020/2398146