Satellite-Driven Monitoring and Prediction of Atmospheric Nitrogen Dioxide in Lithuania

Mina Adel Shokry Fahim* and Jūratė Sužiedelytė Visockienė

Department of Geodesy and Cadastre, Environmental Engineering Faculty, Vilnius Gediminas Technical University, Vilnius, Lithuania
Email: mina-adel-shokry.fahim@vilniustech.lt (M.A.S.F.); juratesuziedelyte-visockiene@vilniustech.lt (J.S.V)
*Corresponding author

Manuscript received February 23, 2025; revised March 19, 2025; accepted April 14, 2025; published July 14, 2025

***Abstract*—Challenges in air pollution control and environmental management continue to evolve, significantly impacting public health and ecosystems. This study estimates surface Nitrogen Dioxide (NO_2) concentrations in Lithuania using an integrated framework. It combines Sentinel-5 Precursor satellite (5P) TROPOspheric Monitoring Instrument (TROPOMI) observations of total NO_2 Vertical Column Densities (VCDs) with various meteorological parameters, such as air temperature, wind speed, and direction. Data were collected across thirteen in situ monitoring stations throughout Lithuania, spanning from January 2020 to December 2023. The analysis involved two modelling approaches: Bagged Trees (BT) and Coarse Tree (CT) algorithms. The BT model outperformed the CT, achieving lower Root Mean Square Error (RMSE) values of 5.12 μg/m³ during validation and 4.90 μg/m³ during testing, compared to 5.73 μg/m³ and 5.58 μg/m³ for the CT model, respectively. The integration of VCDs data and predictive modeling provides valuable indicators for NO_2 concentration trends, stressing the necessity of continuous monitoring efforts, essential for effective air pollution management.**

***Keywords*—satellite monitoring, Nitrogen Dioxide (NO_2), air pollution control, environmental management, atmospheric monitoring, geospatial analysis, air quality assessment, predictive modelling**

I. Introduction

Air pollution still threatens the environment by negatively impacting human health and ecosystems and contributing to climate change. Rapid industrial expansion and increased vehicular traffic have significantly disturbed Earth's atmospheric balance. The disturbance has led to elevated concentrations of harmful gases such as NO_2, with approximately 31.8% of Lithuania's total greenhouse gas emissions derived from the transport sector [1]. Highlighting the urgency of monitoring and mitigating NO_2 levels due to the serious risks to our health, including physical and physiological effects, as well as potential links to cancer and its role in smog, acid rain formation, and respiratory health problems [2]. Historically, the majority of NO_2 levels were measured using a network of monitoring stations located on the ground, which are generally limited in regional extent and tend to miss the air quality status in a regional context. Nevertheless, the advancement of technology on board satellites today facilitates air pollution measurements over large spatial areas and time frames. For example, the Copernicus Sentinel-5 Precursor (Sentinel-5P) has improved environmental geography by enabling atmospheric data collection and evaluation of pollution levels across large areas [3]. These satellite observations offer a robust dataset for applying sophisticated analytical techniques. While there is limited research regionally, the significance of this study lies in its integration of satellite-derived NO_2 Vertical Column Densities (VCDs) with ground-based meteorological variables such as temperature, wind speed, wind direction, wind gust, humidity, and cloud cover. Utilizing advanced predictive algorithms from January 2020 to December 2023 across Lithuania, it assesses the effectiveness of the proposed method for using assembly algorithms BT and regression model CT for forecasting NO_2.

II. Related Study

Vehicles and other combustion-based sources are the main sources of NO_2, which is released in the form of exhaust gases when nitrogen combines with oxygen in the combustion of fuels in automobiles, like vehicles, trucks, and buses. Factories, power plants, and other places that burn fossil fuel are a source of production of NO_2. Residents using natural gas, oil, coal, or wood for heating contribute, especially in cities. Naturally, fires release NO_2 and other nitrogen oxides; storms strike, causing temperatures intense enough to produce NO_2. Therefore, NOXs are also visible in quite substantial quantities in aviation during take-off and landing processes and even in ships and marine while using heavy fuels [4, 5]. On the other hand, it plays a critical role responsible for creating ground-level ozone [6], which destroys plants, especially in cities. Contaminated NO_2 also combines with existing water vapour to yield nitric acid, its primary contribution to the phenomenon of acid rain. As a result, soils and water bodies become acidic, affecting forests, aquatic systems, and biodiversity [7]. Atmospheric deposition of nitrogen compounds involves their transfer from the atmosphere to the Earth's surface, which occurs through various forms like precipitation, dust, and gases [8]. In aquatic systems, when NO_2 gets deposited, it triggers cases of eutrophication, thus promoting the uncontrolled growth of algal blooms, which leads to oxygen supply depletion and threatens the aquatic ecosystem [9]. NO_2 does contribute further towards climate damage as it is not itself a greenhouse gas, but it does, however, help with the process of ozone growth. For agriculture, NO_2 absorption is also a side effect that alters the balance of nutrients in the soils, lowers their productivity, and affects the stability of ecosystems suffering from marked depletion of plant species that are sensitive to nitrogen [10]. The integration of Geographic Information Systems (GIS) with satellite technologies is one of the advances in the domain of air pollution monitoring. This development provided scientists, society and decision-makers with the appropriate mechanisms for monitoring, evaluating, and even managing air pollution in such extensive areas. In contrast to ground-based monitoring stations, satellite-based GIS technologies allow a wide spatial extent, regular temporal resolution, and the ability to study pollutant

doi: 10.18178/ijesd.2025.16.4.1533

levels over inaccessible or difficult-to-reach places. It has become quite a lot easier to record critical pollutants like NO_2 and other gases, the growing access to satellite images like those from NASA's Earth Observation Systems Terra and Aqua satellites, the European Space Agency's (ESA) Sentinel-5P, and the Copernicus Atmospheric Monitoring Service (CAMS). The satellites are equipped with devices such as the TROPOMI or the Ozone Monitoring Instrument (OMI), which allow the determination of atmospheric levels of pollutants by the absorption and scattering of solar radiation by the Earth's atmosphere [11, 12]. While TROPOMI allows the high-resolution data collection of NO_2, CO, SO_2, and other gases [13]. This is particularly important for cities and areas where pollution changes. Such datasets are further integrated into GIS technologies that enable the creation of dynamic pollution spatial and temporal distribution maps for enhanced comprehension of local and regional air quality variations. Spatial interpolation by GIS analysis can estimate in-between pollutant concentrations even for regions where monitoring stations are absent [14, 15]. This becomes important in creating more thorough air quality maps, at least in places with limited monitoring stations. Also, it aims to reveal pollution sources and their transport pathways and assess exposure risk to the affected population. In the event of the COVID-19 pandemic. In the years of global lockdowns, the satellite data monitored noticeable drops in the NO_2 and Particulate Matter (PM) levels over the large cities, which demonstrated the dependence of air quality on human activities. Research conducted in Punjab, Pakistan, examined air quality during the before, during, and post-lockdown phases, evaluating pollutant concentrations such as CO and PM2.5 to comprehend spatial disparities and related health hazards. This shows that the epidemic enhanced air quality, decreasing average CO levels to 0.09 mg/m^3 post-lockdown [16]. The GIS capabilities granted researchers the opportunity to link these with the alterations in transportation system performance, industrial sites' work, and meteorological characteristics, thus having a better insight into the determinants of air pollution [17, 18]. Which advocates for a long-term sustainable policy approach that comprehensively integrates environmental, social, and economic factors [19]. highlight the urgent need for interventions due to escalating air pollution levels, which frequently exceed environmental standards, threatening public health and environmental quality [20].

The use of Machine Learning (ML) and its algorithms has increased significantly in recent years; many studies have been conducted for predictions of air pollutants, such as [21], which have enhanced the precision of air quality forecasts through the use of hybrid deep learning models. As such a study used hyperparameters for Long Short-Term Memory (LSTM), Bidirectional LSTM, and Encoder LSTM models with 500, 500, and 1000 epochs to predict PM2.5 and PM10 levels 5 days ahead, showing superior performance of Encoder LSTM one day ahead over other methods in terms MAE, RMSE, and R^2. proposed five extended features could improve model performance, reducing errors and simplifying the computational complexity of the models. Another study was carried out using NO_2 and O_3 data from 530 and 293 monitoring stations located across Italy. In order to generate high spatial resolution NO_2 and O_3 daily concentration over the years 2013–2015, an integrated approach that integrated a Chemical Transport Model (CTM) with ML random forest algorithm, executed by the Flexible Air quality Regional Model (FARM) as a predictor when used with other spatial-temporal data, such as population, land-use, surface greenness, and road networks. The ML improvements helped mitigate the underestimation of NO_2 concentration, demonstrating that in 2014, RMSE (µg/m^3) for NO_2 and O_3 were 11.7 and 14.2, respectively [22]. In China [23], from May 2013 to April 2014, by using on year dataset of metrological variables and tropospheric NO_2 columns from the OMI to estimate NO_2 for ground level using a geographically and temporally weighted regression model was introduced to treat the spatiotemporal non-stationarities-between-tropospheric-columnar-and ground level NO_2, the assessment of cross-validation over ordinary least squares, geographically weighted regression, temporally weighted regression, geographically and temporally weighted regression shows that geographically and temporally weighted regression best performance with R^2 by 0.6.

Four ML algorithms were used for East Asian air quality assessment: random forest, support vector machine, extreme gradient boosting, and light gradient boosting to estimate NO_2 and O_3 concentration [24]. These algorithms have been compared with multiple linear regression using a variety of data, including TROPOMI, metrological, and land use variables, and they have been validated using three different cross-validation techniques: random cross-validation, spatial cross-validation, and temporal cross-validation. The model for NO_2 had R^2 of 0.63 to 0.7 and an RMSE of 4.76 to 5.28, while the model for O_3 had R^2 of 0.65 to 0.78 and an RMSE of 13.96 to 11.11. The light gradient boosting random cross-validation produced the best results. Across Europe, a study effectively utilizes a Satellite and ML-based Estimation of Surface air quality at a High-resolution model, combining satellite data and ML (XGBoost), to estimate daily surface NO_2 concentrations at a high spatial resolution of 1 km [25]. The study demonstrates good predictive accuracy, evidenced by a mean absolute error of 7.77 µg/m^3 and a Spearman rank correlation of 0.66. This demonstrates the reliability of using TROPOMI data, even with a spatial resolution of 7.5 × 3.5 km. Furthermore, it demonstrates the possibility of deriving higher-resolution spatial maps of surface NO_2 (such as 1 km) by synergistically utilizing the higher spatial resolution of representative proxy features such as Visible Infrared Imaging Radiometer Suite (VIIRS) night lights.

III. Materials and Methods

The methodology of this study involved several structured steps to acquire, process, and analyze data for NO_2 concentrations and meteorological variables from various monitoring stations in Lithuania (see Fig. 1).

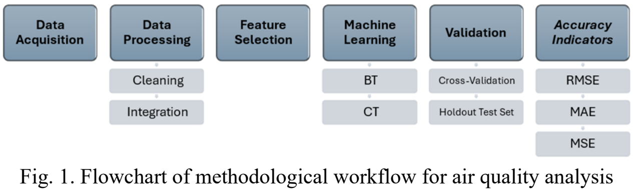

Fig. 1. Flowchart of methodological workflow for air quality analysis (source: the authors).

A. Data Acquisition and Preprocessing

The NO_2 dataset utilized in this research covers the span of

2020 to 2023 and includes observations from the TROPOMI sensor on the Sentinel-5P satellite. These data were obtained from the Copernicus program using Google Earth Engine (GEE) API. Approximately 29,300 Earth observations were stored on Google's cloud platform, as illustrated in Fig. 2. A third-degree polynomial was fitted across Lithuania to analyze changes in NO_2 concentration over the dedicated years at 13 distributed monitoring stations. In total, this analysis utilized approximately 455,800 data observations from monitoring stations over the four-year timeframe. Table 1 presents the statistics of these measurements after the cleaning phase. The data for monitoring station observations were obtained from the European Environment Agency (EEA) [26]. Also, meteorological data relevant to the cities were obtained from the Lithuanian Hydrometeorological Service.

Table 1. Statistical of NO_2 concentrations (µg/m³) post-cleaning for monitoring stations

Station Code	mean	median	std	min	max	Range
LT00001	11.16	9.6	6.25	2.87	37.28	34.41
LT00002	9.37	7.8	5.60	2.87	37.3	34.43
LT00003	21.74	21.405	8.12	3.3	37.5	34.2
LT00004	12.63	11.3	7.01	2.87	37.5	34.63
LT00012	11.14	9.6	6.04	2.87	36.5	33.63
LT00022	18.06	17.4	7.08	3.4	37.1	33.7
LT00023	6.64	5.5	3.97	2.87	36.52	33.65
LT00031	12.56	10.7	7.18	2.87	37.3	34.43
LT00033	17.74	16.8	7.99	2.87	37.48	34.61
LT00041	17.16	15.87	8.54	2.9	37.5	34.6
LT00043	7.52	6.5	4.13	2.87	33.46	30.59
LT00044	6.77	5.5	4.44	2.87	33.5	30.63
LT00052	4.41	3.8	1.76	2.87	13.19	10.32

The study includes a diverse range of meteorological information, including air temperature, feels-like temperature, wind speed, wind gust, wind direction, cloud cover, sea level pressure, relative humidity, and precipitation. Subjected to TROPOMI data for various monitoring station coordinates that have been extracted with respect to the nearest pixel throughout the dedicated analysis period. We modelled it using a third-degree polynomial curve to project trends and patterns over the whole period "Fig. 2".

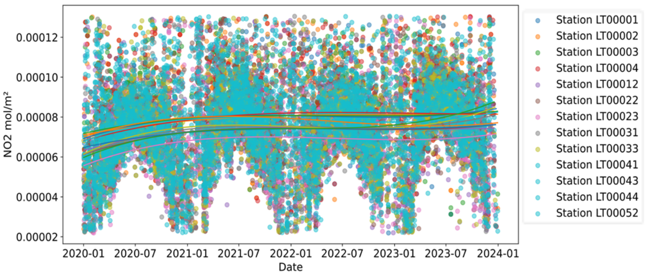

Fig. 2. Observations of TROPOMI with respect to the location of each station and Representation of Third-Degree Polynomial Over Dedicated Period (source: the authors).

The data cleaning process is applied to a data frame. Initially, rows were filtered out where either the 'Validity' or 'Verification' fields didn't meet. Next, by selecting non-positive values for NO_2 monitoring station observations and VCD_s were removed to ensure data quality. The sources of all datasets used are listed in Table 2. Following this, we implemented an interquartile range (IQR) filtering for VCD_s and station observation values. This involved calculating the first and third quartiles (Q1 and Q3) and the IQR for these values and then excluding any values that fall more than 1.5 times IQR below Q1 or above Q3 to mitigate the effect of outliers and refine the dataset for further analysis.

B. Feature Selection

The study focused on data from thirteen monitoring stations located across Lithuania. NO_2 concentrations were fetched for the period between January 1, 2020, and December 31, 2023. The TROPOMI sensor on the Sentinel-5P satellite was accessed through the Copernicus program to gather the total VCDs. The deployment of the Sentinel-5P satellite is important for monitoring and forecasting ground-level NO_2 concentrations [25].

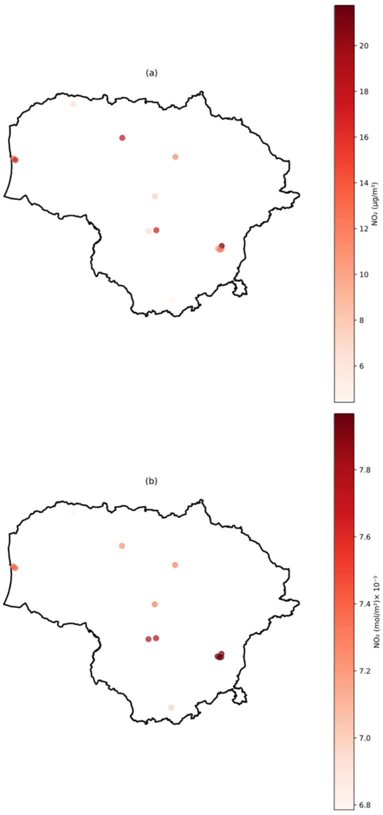

Fig. 3. Concentrations NO_2: (a) Monitoring Station Mean Values and (b) TROPOMI Sentinel-5P Derived Observations Over Four Years (source: the authors).

Fig. 3a shows the distribution of NO_2 concentrations across various monitoring stations plotted and outlined with the Lithuanian border. The colour scale ranges from lower concentrations to higher concentrations. Each dot represents a monitoring station, with the colour indicating the mean concentration recorded at that location overall for the total period, while Fig. 3b presents a similar distribution variance but highlights the data obtained from satellite observations. The purpose of these figures is to show the similarity and visually compare both data sets, focusing on the valuable use of satellite data for estimating surface NO_2 concentrations.

A custom Python script was employed to retrieve the NO_2 VCDs data for each of the monitoring stations through a point-based query. The extracted data was structured into a pandas data frame. The acquired satellite-derived VCDs and the levels of monitoring station NO_2 concentrations were employed to the nearest point and then merged with the other meteorological data regarding each city. The final dataset constituted retrieved values for NO_2 concentrations. The data was employed in time-series analyses aimed at understanding predictors, as well as machine-learning model formulation.

Table 2. Input variables used to estimate surface NO_2

Dataset	Variable	Type	Data source
Monitoring stations	surface NO_2 concentrations	Numeric	EEA
TROPOMI	Total vertical column of NO_2	Numeric	Sentinel-5p
Metrological conditions	air temperature, feels-like temperature, wind speed, wind gust, wind direction, cloud cover, sea level pressure, relative humidity, and precipitation	Numeric	Lithuanian Hydrometeorological Service
Station Code		Numeric	
City Id		Numeric	
date time serial		Numeric	

A. Machine Learning

Once the data has been preprocessed, the following procedures to prepare it for testing and training. In deep learning, experiments are conducted using various techniques to process data and create models. In this paper, we aim to utilize BT and CT algorithms as they are robust and capable of capturing complex nonlinear relationships within datasets. Specifically, ML techniques were employed to efficiently predict NO_2 concentrations. The BT Algorithm operates by employing an ensemble method in which multiple decision trees are created based on different parts of the dataset, and their outputs are averaged. BT is fairly effective in solving variance and overfitting problems [27, 28], which are common barriers in the process of analyzing the information that relates to the environmental aspects, explaining that the BT algorithm has its origins in the work represented in Eq. (1) [28]. The CT Algorithm employs large splits, which effectively reduce the complexity of decision trees. This method is useful in developing information that is more understanding about the data, although it is more appropriate for the first examination and investigation of a pattern within the data rather than predicting it narrowly [29].

$$f(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{1}$$

where $f(x)$ represents the averaged prediction from B decision trees, where $Tb(x)$ is the prediction from the b^{th} tree in the ensemble

B. Accuracy Indicators

Upon completion of training and model creation, 5-fold Cross-validation was employed to evaluate the model's performance. A key metric for model evaluation is the mean squared error (MSE), which is determined for each fold, as illustrated in Eq. (2). The overall test MSE is calculated by averaging the individual MSE values computed across all folds, as shown in Eq. (3) [30, 31] This approach involves splitting the dataset into five sets. In each cycle, one set serves as the validation set, while the remaining four sets are used for training to ensure that each data point is used once for both training and validation, enhancing the model's reliability.

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

where n is the number of observations in the fold, y_i is the actual value of the ith observation, and $\hat{y}_i$ is the predicted value by the model for the ith observation.

$$\text{Test MSE} = \frac{1}{k}\sum_{i=1}^{k}\text{MSE}_i \tag{3}$$

where k represents the number of folds and the MSE of the ith fold.

Mean absolute error (MAE) has been utilized in the fields of air quality management as well as atmospheric studies as it is a concept that allows for easy and strong comparison of model predictions against the observations as regarding Eq. (4). which is essential in determining how accurately the concentration of pollutants can be forecasted [32, 33].

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

where n is the number of observations, y_i is the actual value of the ith observation, and $\hat{y}_i$ is the predicted value by the model for the ith observation.

The effectiveness of the models may also be gauged by the residual plots, by the validation of Predicted vs. Actual plots, and by the comparative analysis of the BT and CT algorithms. Illustrations display how well the model forecasts correspond to the observed variables. During validation, the Predicted vs Actual plots visually evaluate the precision of the models. In Fig. 4, each algorithm applies to new data by providing scatter plots of predicted versus true values [34].

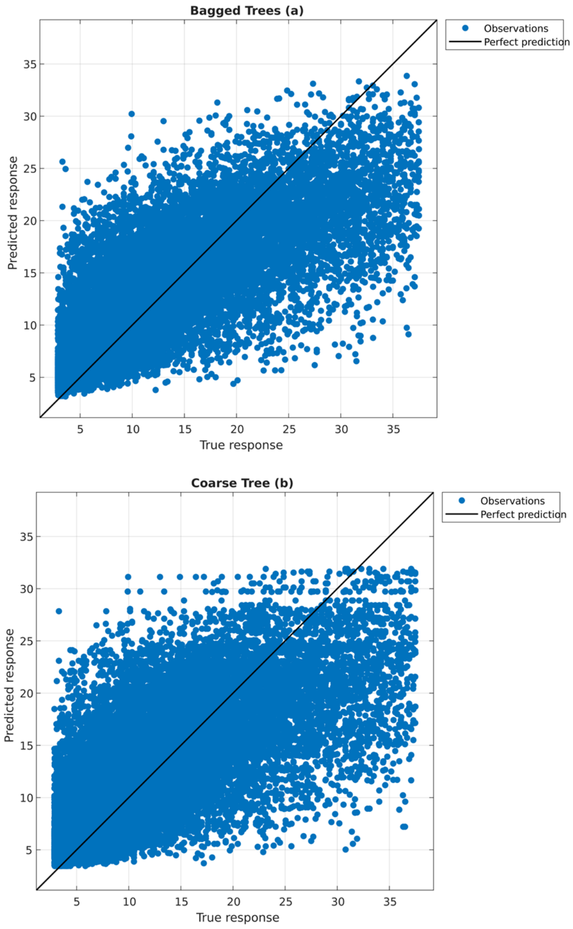

Fig. 4. Predicted vs. Actual Plot for Validation Phase – (a) BT and (b) CT Models (source: the authors).

Out of the total data, 10% was set aside for testing purposes. The separate set facilitates the unbiased evaluation of the model's performance in practical cases that lie outside the training data in the testing phase cycle. The RMSE, MSE, and MAE [35] The evaluation of accuracy and effectiveness for the developed model. In the testing stage, the researcher assesses the presence of the systematic components in the prediction error by plotting the residuals against the predicted values. A fitted model displays residuals scattered randomly around the horizontal axis, indicating uniform variance. Fig. 5 was used to identify model behaviour issues and determine whether residuals exhibit non-random patterns for BT and CT.

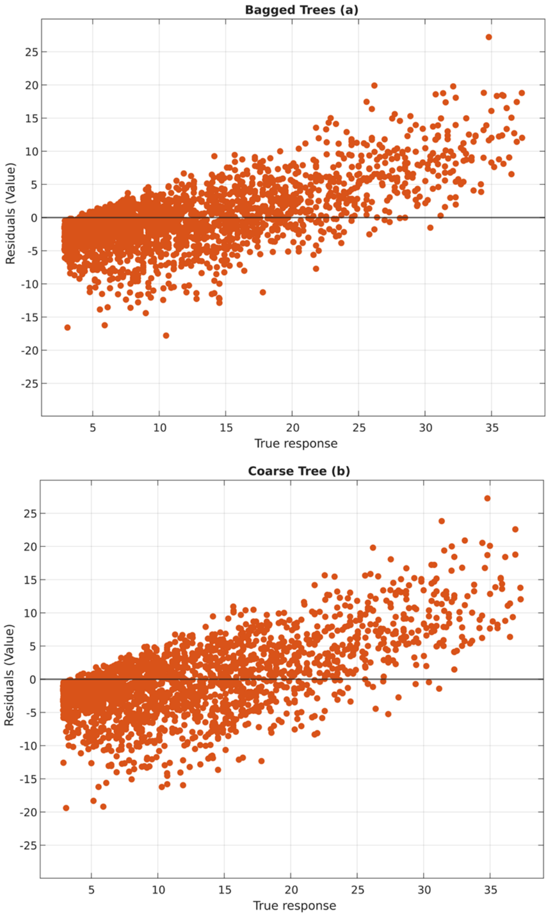

Fig. 5. Residuals Plots for Testing Phase - (a) BT and (b) CT Models (source: the authors).

IV. Result

The BT model shows a validation RMSE of 5.12 µg/m³ and a test RMSE of 4.90 µg/m³, with MSE values of 26.18 (µg/m³)² (validation) and 23.97 (µg/m³)² (test), and MAE values of 3.71 µg/m³ (validation) and 3.55 µg/m³ (test), indicating its effectiveness in handling the dataset using 30 learners. In comparison, the CT model has a slightly lower accuracy with a validation RMSE of 5.73 µg/m³ and a test RMSE of 5.58 µg/m³, along with MSE values of 32.86 (µg/m³)² (validation) and 31.14 (µg/m³)² (test) and MAE values of 4.20 µg/m³ (validation) and 4.15 µg/m³ (test) as referred Table 3. Despite its lower accuracy, the CT model has a faster prediction speed, and the outperformed BT emphasizes its potential to be used in further analysis.

Table 3. Performances of models

Model	RMSE	MSE	MAE	RMSE	MSE	MAE
	Validation			Test		
BT	5.12	26.18	3.71	4.90	23.97	3.55
CT	5.73	32.86	4.20	5.58	31.14	4.15

The use of satellite data had a positive impact on the accuracy of NO_2 concentration prediction at each monitoring location in Lithuania. The integration of TROPOMI data enhanced the examination of trends in the areas of interest. Between January 2020 and December 2023, for TROPOMI sensor data, VCDs peaked between June and August for the majority of stations, suggesting a decrease in air quality. The upward trend in NO_2 levels for the total period is displayed by the third-degree polynomial on the graph throughout the post-cleaning phase. The trend of increasing NO_2 concentrations in Lithuania is leading to a need for further research and action in the domain of air quality.

V. Discussion

The results demonstrate that the BT ensemble model provides a better predictive performance than the single decision CT model. Across all error metrics RMSE, MSE, and MAE, the BT model achieved lower errors, underscoring the value of ensemble techniques in modelling NO_2. The finding aligns with the well-known advantage of ensemble learning by aggregating many decision trees, each trained on bootstrap samples. The model reduces variance and avoids overfitting noise. In practical terms, our BT model's predictions of ground-level NO_2 were considerably more accurate and stable than the CT model's, confirming that the integration of multiple learners yields a better predictor. Similar observations have been reported in other air quality studies [36, 37]. The integration of satellite-derived TROPOMI NO_2 column data with ground-based meteorological inputs led to improved prediction quality. The satellite Tropospheric VCDs of NO_2 from Sentinel-5P proved to be highly informative predictors when combined with local meteorological variables (e.g. temperature, wind, etc.), allowing the BT model to learn spatiotemporal patterns that a ground-only model might miss. Our findings are in line with other research that fuses remote sensing with in-situ data. For instance, [37] combined Sentinel-5P NO_2 observations with ground meteorological measurements to estimate surface NO_2 and found that advanced machine learning methods dramatically lowered prediction error (achieving RMSE ~2.9 µg/m³) compared to models using satellite data alone.

Temporal patterns in Lithuania revealed a peak during the summer months. Typically, urban NO_2 concentrations peak in winter due to increased heating emissions and weaker atmospheric dispersion and dip in summer. However, the data of VCDs showed higher NO_2 in summertime, which we attribute to increased photochemical activity and transport-related emissions in the warm season. Strong sunlight and higher temperatures in summer catalyze photochemical reactions that produce ozone and other oxidants; under certain conditions, these reactions can also elevate NO_2 levels through the photo-stationary equilibrium between NO, NO_2, and O_3. Empirical studies in Europe have noted that summer NO_2 episodes often coincide with ozone pollution events during hot, stagnant weather when ample sunlight drives photochemical NO_x oxidation [37].

VI. Conclusion

The research combined various data sources, including TROPOMI sensor data from the Sentinel-5P satellite, to examine NO_2 concentrations at thirteen monitoring stations in Lithuania from 2020 to 2023. Meteorological parameters were also incorporated. The results of the modelling techniques, BT and CT algorithms, demonstrated an approach to predicting NO_2 levels. Within the validation and testing phases, the BT model exhibited superior accuracy compared to the CT model, achieving RMSE, MSE, and MAE values of 4.9, 23.97, and 3.55, respectively. This highlights the effectiveness of ensemble methods in addressing environmental data challenges. Statistical analysis indicated an increase in NO_2 VCDs throughout the study period, suggesting a pressing need for continuous monitoring and intervention strategies to mitigate air pollution. The utilization of TROPOMI satellite data proved valuable in enhancing the predictive capability of the ground-based monitoring system, underscoring the importance of leveraging satellite technology in atmospheric research.

VII. Recommendation

While this study focused on NO_2, the approach of satellite-driven ML modelling can be extended to other pollutants and can benefit from emerging machine-learning techniques. Sentinel-5P's payload TROPOMI measures several atmospheric constituents, including ozone, sulfur dioxide (SO_2), carbon monoxide (CO), formaldehyde, and aerosol optical depth. There is an opportunity to expand the methodology to these pollutants. For example, ground-level ozone could be predicted by blending Sentinel-5P ozone column data with meteorological factors, temperature, sunlight, and VOC indicators in an ensemble model. One challenge is that ozone formation is highly non-linear, but ML models, especially ensemble models, can potentially capture such chemistry-climate interactions. Another extension is towards particulate matter $PM_2.5$ and PM_{10}. Although Sentinel-5P does not measure PM directly, it provides aerosol indices and precursors like NO_2 and SO_2, which contribute to secondary aerosol formation. Combining these with other satellite data and meteorology in an ML framework could allow the estimation of surface PM levels.

Limitation

Overall, the primary disadvantage of the research is the insufficient verification for some monitoring station observations, which resulted in several data points having to be discarded. Furthermore, the absence of metrological parameters data for some cities led to their being dropped from the analysis. Although the methodology approach's potential in NO_2 predictions has been successful, there is still room for improvement by adding more indicators. As such, further refinements can include the traffic counts, pollution caused by industries and pollution crossing borders to improve the estimation accuracy.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

The first author's responsibilities included data collection and analysis, leading to the creation of the initial draft. The second author, however, took the lead in defining the research scope and guiding the paper to its final completion.

References

[1] Greenhouse gas emissions from transport are falling too slowly State control of the Republic of Lithuania. (Sep. 20, 2024). [Online]. Available: https://www.valstybeskontrole.lt/EN/Post/17837/greenhouse-gas-emissions-from-transport-are-falling-too-slowly

[2] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, "Environmental and health impacts of air pollution: A review," *Frontiers Media S.A*, Feb. 20, 2020. doi: 10.3389/fpubh.2020.00014

[3] ESA - Sentinel-5P brings air pollution into focus. (Sep. 03, 2024). [Online]. Available: https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-5P/Sentinel-5P_brings_air_pollution_into_focus

[4] B. Kuang *et al.*, "Chemical characterization, formation mechanisms and source apportionment of PM2.5 in north Zhejiang Province: The importance of secondary formation and vehicle emission," *Science of The Total Environment*, vol. 851, p. 158206, Dec. 2022. doi: 10.1016/J.SCITOTENV.2022.158206

[5] A. Fayyazbakhsh *et al.*, "Engine emissions with air pollutants and greenhouse gases and their control technologies," *J Clean Prod*, vol. 376, p. 134260, Nov. 2022. doi: 10.1016/J.JCLEPRO.2022.134260.

[6] Ground-level Ozone Basics | US EPA. (Accessed: Oct. 20, 2024). [Online]. Available: https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics

[7] C. Jia *et al.*, "Pollution characteristics and potential sources of nitrous acid (HONO) in early autumn 2018 of Beijing," *Science of the Total Environment*, vol. 735, Sep. 2020. doi: 10.1016/j.scitotenv.2020.139317

[8] C. H. Zhang *et al.*, "Atmospheric nitrogen deposition and its responses to anthropogenic emissions in a global hotspot region," *Atmos Res*, vol. 248, Jan. 2021. doi: 10.1016/j.atmosres.2020.105137

[9] Z. Yu, X. Sun, L. Yan, S. Yu, Y. Li, and H. Jin, "Analysis of the water quality status and its historical evolution trend in the mainstream and major tributaries of the yellow river basin," *Water (Switzerland)*, vol. 16, no. 17, Sep. 2024. doi: 10.3390/w16172413

[10] J. Weber, J. Keeble, N. L. Abraham, D. J. Beerling, and M. V. Martin, "Global agricultural N_2O emission reduction strategies deliver climate benefits with minimal impact on stratospheric O3 recovery," *NPJ Clim Atmos Sci*, vol. 7, no. 1, Dec. 2024. doi: 10.1038/S41612-024-00678-2

[11] D. Stratoulias, N. Nuthammachot, R. Dejchanchaiwong, P. Tekasakul, and G. R. Carmichael, "Recent developments in satellite remote sensing for air pollution surveillance in support of sustainable development goals," *Remote Sens (Basel)*, vol. 16, no. 16, Aug. 2024. doi: 10.3390/rs16162932

[12] M. Vîrghileanu, I. Săvulescu, B. A. Mihai, C. Nistor, and R. Dobre, "Nitrogen dioxide (NO_2) pollution monitoring with sentinel-5p satellite imagery over europe during the coronavirus pandemic outbreak," *Remote Sens (Basel)*, vol. 12, no. 21, pp. 1–29, Nov. 2020. doi: 10.3390/rs12213575

[13] J. Van Geffen *et al.*, "S5P TROPOMI NO_2 slant column retrieval: Method, stability, uncertainties and comparisons with OMI," *Atmos Meas Tech*, vol. 13, no. 3, pp. 1315–1335, Mar. 2020. doi: 10.5194/AMT-13-1315-2020

[14] Y. Bezyk, I. Sówka, M. Górka, and J. Blachowski, "Gis-based approach to spatio-temporal interpolation of atmospheric co2 concentrations in limited monitoring dataset," *Atmosphere (Basel)*, vol. 12, no. 3, Mar. 2021. doi: 10.3390/atmos12030384

[15] L. Matejicek, "Spatial modelling of air pollution in urban areas with GIS: A case study on integrated database development," *Advances in Geosciences*, vol. 4, pp. 63–68, 2005.

[16] M. Nasar-u-Minallah, N. Parveen, Bushra, and M. Jabbar, "Assessing air quality dynamics in Punjab, Pakistan: Pre, during, and post COVID-19 lockdown and evaluating strategies for mitigating," *GeoJournal*, vol. 89, no. 4, Aug. 2024. doi: 10.1007/s10708-024-11132-4

[17] NASA Helps Map Impact of COVID-19 Lockdowns on Harmful Air Pollution - NASA. (Oct. 23, 2024). [Online]. Available: https://www.nasa.gov/missions/aqua/nasa-helps-map-impact-of-covid-19-lockdowns-on-harmful-air-pollution/

[18] M. J. Cooper *et al.*, "Global fine-scale changes in ambient NO_2 during COVID-19 lockdowns," *Nature*, vol. 601, 2022. doi: 10.1038/s41586-021-04229-0

[19] M. Nasar-U-Minallah, M. Zainab, and M. Jabbar, "Exploring mitigation strategies for smog crisis in Lahore: A review for environmental health, and policy implications," *Environ Monit Assess*, vol. 196, no. 12, p. 1269, Nov. 2024. doi: 10.1007/s10661-024-13336-0

[20] M. Nasar-u-Minallah, M. Jabbar, S. Zia, and N. Perveen, "Assessing and anticipating environmental challenges in Lahore, Pakistan: Future implications of air pollution on sustainable development and

environmental governance," *Environ Monit Assess*, vol. 196, no. 9, Sep. 2024. doi: 10.1007/s10661-024-12925-3.

[21] T. P. T. Quynh, T. N. Viet, H. D. Thi, and K. H. Manh, "Enhancing air quality prediction accuracy using hybrid deep learning," *International Journal of Environmental Science and Development*, vol. 14, no. 2, pp. 155–159, Apr. 2023. doi: 10.18178/ijesd.2023.14.2.1428.

[22] C. Silibello *et al.*, "Spatial-temporal prediction of ambient nitrogen dioxide and ozone levels over Italy using a Random Forest model for population exposure assessment," *Air Quality Atmosphere & Health*. doi: 10.1007/s11869-021-00981-4/Published.

[23] K. Qin *et al.*, "Estimating ground level NO_2 concentrations over central-eastern China using a satellite-based geographically and temporally weighted regression model," *Remote Sens (Basel)*, vol. 9, no. 9, Sep. 2017. doi: 10.3390/rs9090950.

[24] Y. Kang *et al.*, "Estimation of surface-level NO_2 and O_3 concentrations using TROPOMI data and machine learning over East Asia," *Environmental Pollution*, vol. 288, Nov. 2021. doi: 10.1016/j.envpol.2021.117711.

[25] [S. Shetty, P. Schneider, K. Stebel, P. David Hamer, A. Kylling, and T. Koren Berntsen, "Estimating surface NO_2 concentrations over Europe using Sentinel-5P TROPOMI observations and machine learning," *Remote Sens Environ*, vol. 312, Oct. 2024. doi: 10.1016/j.rse.2024.114321.

[26] European Environment Agency's home page. (Mar. 10, 2024). [Online]. Available: https://www.eea.europa.eu/en

[27] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Mach Learn*, vol. 51, no. 2, pp. 181–207, May 2003. doi: 10.1023/A:1022859003006/METRICS.

[28] [L. Breiman, "Bagging predictors," *Mach Learn*, vol. 24, no. 2, pp. 123–140, 1996. doi: 10.1007/BF00058655/METRICS.

[29] L. A. Breslow and D. W. Aha, "Simplifying decision trees: A survey," *Knowl Eng Rev*, vol. 12, no. 01, pp. 1–40, 1997. doi: 10.1017/S0269888997000015.

[30] S. Bates, T. Hastie, and R. Tibshirani, "Cross-validation: what does it estimate and how well does it do it?" *Journal of the American Statistical Association*, 2022.

[31] Stanford CS Theory. (Sep. 10, 2024). [Online]. Available: https://theory.stanford.edu/main/index.shtml

[32] L. Lin *et al.*, "Estimating $PM_{2.5}$ concentrations using the machine learning RF-XGBoost Model in Guanzhong urban agglomeration, China," *Remote Sens (Basel)*, vol. 14, no. 20, Oct. 2022. doi: 10.3390/rs14205239.

[33] L. Chen, B. Han, X. Wang, J. Zhao, W. Yang, and Z. Yang, "Machine learning methods in weather and climate applications: A survey," *Applied Sciences*, Oct. 30, 2023. doi: 10.20944/preprints202309.1764.v2.

[34] A. T. Tredennick, G. Hooker, S. P. Ellner, and P. B. Adler, "A practical guide to selecting models for exploration, inference, and prediction in ecology," *Ecology*, 2021. doi: 10.1002/ecy.3336.

[35] X. Wen, M. Jaxa-Rozen, and E. Trutnevyte, "Accuracy indicators for evaluating the retrospective performance of energy system models," *Appl Energy*, vol. 325, p. 119906, Nov. 2022. doi: 10.1016/J.APENERGY.2022.119906.

[36] D. Triana and S. Osowski, "Bagging and boosting techniques in prediction of particulate matters," *Bulletin of the Polish Academy of Sciences Technical Sciences*, vol. 68, no. 5, 2020. doi: 10.24425/bpasts.2020.134659.

[37] J. R. Cedeno Jimenez, A. de J. Pugliese Viloria, and M. A. Brovelli, "Estimating daily NO_2 ground level concentrations using sentinel-5P and ground sensor meteorological measurements," *ISPRS Int J Geoinf*, vol. 12, no. 3, Mar. 2023. doi: 10.3390/ijgi12030107.