VERİ BİLİMİ DERGİSİ

www.dergipark.gov.tr/veri

An Analysis on Environmental Justice and Air Quality Using Machine Learning Techniques

Görkem DEMİRCAN[1*], Gülsüm Çiğdem ÇAVDAROĞLU[2]

[1, 2]*Isik University, Faculty of Economics, Administrative and Social Sciences, İstanbul*

Abstract

This study examines air quality dynamics across countries using machine learning with a focus on environmental justice. Random Forest, Decision Tree, XGBoost, and Adaboost algorithms were applied for a 10-year air pollution forecast. XGBoost showed the best performance. Increases in pollutant levels are expected in Bhutan and North Korea, while improvements may occur in India, Pakistan, and Nepal. Significant air quality changes are projected in Laos, Indonesia, and North Korea. The study highlights inequalities in pollution exposure and emphasizes the need for targeted interventions.

Keywords: Air Quality Prediction, CO Level Prediction, Machine Learning

Makale Bilgisi

Başvuru: 17/11/2024
Kabul: 07/07/2025

Makine Öğrenimi Teknikleri ile Çevresel Adalet ve Hava Kalitesi Üzerine Bir Analiz

Özet

Bu çalışma, çevresel adalet odağında makine öğrenmesi yöntemleriyle farklı ülkelerde hava kalitesi dinamiklerini incelemektedir. Random Forest, Karar Ağacı, XGBoost ve Adaboost algoritmaları kullanılarak 10 yıllık hava kirliliği tahmini yapılmıştır. En iyi performans XGBoost modelinde görülmüştür. Bhutan ve Kuzey Kore'de kirletici seviyelerinin artacağı, Hindistan, Pakistan ve Nepal'de ise iyileşmeler yaşanabileceği öngörülmüştür. Laos, Endonezya ve Kuzey Kore'de önemli hava kalitesi değişiklikleri beklenmektedir. Çalışma, kirlilikteki eşitsizliklere dikkat çekerek hedefe yönelik müdahalelerin gerekliliğini vurgular.

Anahtar Kelimeler: *Hava Kalitesi Tahmini, CO Seviyesi Tahmini, Makine Öğrenmesi.*

* Corresponding e-mail: 20MISY1044@isik.edu.tr

1 Introduction

Rapid urbanization, industrial growth, and global environmental challenges have underscored the critical importance of air quality studies, given their profound impact on ecological well-being and public health. As cities expand and industries proliferate, pollutant emissions have become a major contributor to the degradation of natural environments and the deterioration of human living conditions. This study employs machine learning (computer algorithms that learn patterns from data) and advanced data analysis techniques to examine air quality, bridging interdisciplinary gaps by analyzing historical trends and developing robust predictive models.

In recent decades, the discourse on climate change has evolved significantly. The United Nations Framework Convention on Climate Change has broadened the definition of climate change to include human-induced factors, linking these factors closely with the escalating global climate crisis. The climate crisis is characterized not only by the gradual increase in global temperatures but also by the intensification of extreme weather events, erratic precipitation patterns, prolonged droughts, and other disruptive environmental phenomena. These developments are interwoven with air quality issues, as pollutants contribute both to the warming of the atmosphere and to a cascade of adverse health effects.

The urgency to address these interlinked challenges has never been greater. Global warming, driven primarily by the emission of greenhouse gases, accelerates the melting of polar ice and disrupts established climatic patterns. This, in turn, exacerbates the frequency and severity of meteorological events, such as hurricanes and heatwaves, which pose significant risks to human populations and natural ecosystems. The rapid deterioration of climate conditions is now seen as a critical threat to sustainable development, requiring an integrated approach that combines environmental science, public policy, and technological innovation.

Recognizing that the impacts of climate change are not uniform, this study also emphasizes the concept of environmental injustice. Vulnerable populations and regions—often lacking the resources to mitigate or adapt to these changes—are disproportionately affected by both air pollution and the broader effects of the climate crisis. Identifying these disparities is crucial for formulating fair and effective environmental policies. Equitable solutions require an understanding of not only the scientific and technical aspects of air quality but also the socio-economic dimensions of environmental degradation.

Against this backdrop, our research provides a comprehensive overview of air pollution through the analysis of data from multiple countries. We have developed and applied several machine learning models—including Random Forest, Decision Tree, XGBoost, and Adaboost—to forecast future air quality trends over a decade. By examining spatial and temporal variations in pollutant levels, the study offers insights into how different regions might experience changes in air quality. Additionally, our work underscores the importance of advanced data visualization techniques, which transform complex datasets into accessible information for policymakers and public health officials.

This study not only contributes to the academic discourse on climate change and air quality but also aims to support informed decision-making. By integrating technological innovations with environmental research, we strive to provide a framework for anticipating future challenges and devising strategies that address both the scientific and societal dimensions of the climate crisis.

2 Literature review

The application of machine learning algorithms for air quality forecasting has gained significant momentum in recent years, with researchers employing diverse methodological approaches to enhance prediction accuracy. Ensemble methods (techniques that combine multiple algorithms for better predictions) have emerged as particularly effective solutions, with gradient boosting variants (algorithms that build models sequentially to correct previous errors) consistently demonstrating superior performance across multiple studies [27], [23]. LightGBM has been identified as especially suitable for high-dimensional datasets due to its histogram-based algorithm (a method that groups data into bins for faster processing) and parallel learning capabilities (ability to process data simultaneously across multiple processors) [27], while CatBoost has shown exceptional results with correlation coefficients reaching 0.9998 in some applications [23].

Traditional regression approaches (statistical methods that model relationships between variables) continue to play important roles in air quality modeling. Multiple linear regression without feature selection has proven competitive with more complex algorithms in certain contexts, particularly for particulate matter prediction [16]. However, the integration of advanced techniques such as support vector machines and neural networks has expanded the analytical toolkit available to researchers [5], with convolutional neural networks showing promise for spatial distribution modeling (analyzing how pollutants vary across geographical areas) of pollutants [25].

Time series analysis (methods for analyzing data points collected over time) remains fundamental to air quality prediction, with researchers utilizing various temporal modeling approaches to capture seasonal patterns and long-term trends [2, 5]. The incorporation of meteorological parameters has proven crucial, with temperature, humidity, and atmospheric pressure emerging as key predictive variables [17]. Long Short-Term Memory networks have demonstrated effectiveness for temporal sequence modeling, achieving goodness-of-fit scores exceeding 90% in air quality index predictions [17].

Spatial considerations have gained increasing attention, with studies revealing significant geographical variations in pollutant concentrations. The identification of industrial areas as primary sources of sulfur dioxide contamination exemplifies the importance of spatial context in environmental modeling [2]. Advanced approaches incorporating building height, topography, and emission sources have enhanced the accuracy of spatial pollution distribution models [25].

Dataset imbalance (when some categories in the data have significantly fewer examples than others) has emerged as a critical challenge in air quality classification tasks. The implementation of Synthetic Minority Oversampling Technique (SMOTE — a method to balance datasets by creating synthetic examples) has demonstrated significant improvements in model performance, with accuracy gains of up to 20% observed in balanced datasets [10]. Feature selection (the process of choosing the most relevant variables for prediction) and correlation analysis have proven essential for identifying the most influential parameters, with studies typically focusing on 10–12 key variables from larger meteorological and pollutant datasets [17, 1].

Cross-validation methodologies (techniques to test model performance on unseen data), particularly 10-fold approaches, have become standard practice for model validation, ensuring generalizability across different temporal periods and geographical locations [25]. The integration of both historical and forecasted meteorological data has enhanced model robustness, providing more

comprehensive input features for prediction algorithms [17].

Comparative analyses reveal distinct performance characteristics across different algorithmic approaches. While ensemble methods generally outperform individual algorithms, the optimal choice depends significantly on dataset characteristics and prediction objectives [16, 11]. Root Mean Square Error values typically range from 0.1 to 8.0 across studies, with correlation coefficients (statistical measures of the strength of linear relationships between variables, ranging from −1 to +1) commonly exceeding 0.85 for well-performing models [16, 10, 25].

Stack models (ensemble techniques that combine predictions from multiple different algorithms) and ensemble approaches have consistently demonstrated superior performance by mitigating individual model limitations and reducing overfitting tendencies [11, 1]. The combination of varied algorithms through stacking has achieved accuracy scores approaching 97% in air quality grade classification tasks, suggesting the effectiveness of hybrid methodological approaches [1].

Despite extensive research on prediction methodologies, limited attention has been devoted to environmental justice implications of air quality disparities. The unequal distribution of air pollution exposure across diverse populations and geographical regions represents a critical research gap requiring urgent attention. Global air quality patterns reveal significant disparities that extend beyond local impact, influencing ecosystems, climate patterns, and planetary health.

The integration of machine learning capabilities with environmental justice analysis offers promising opportunities for identifying patterns of inequality and developing targeted intervention strategies. This interdisciplinary approach allows a more comprehensive understanding of air quality disparities while providing predictive capabilities for policy development and resource allocation decisions.

3 Methodology

In the presented study, a dataset containing air quality values for different countries was utilized. The data set covers the years 2003-2018 and was acquired from NASA's Atmospheric Science Data Center (https://sedac.ciesin.columbia.edu/data/set/aqdh-country-trends-major-air-pollutants-2003-2018). This repository comprises a rich collection of observational data, capturing key atmospheric parameters crucial for our analyses. Specifically, the dataset was developed to provide public-health-focused air quality indicators, quantifying trends in exposure to major air pollutants across over 200 countries. These pollutants include:

- ✓ Particulate matter (PM): Tiny solid or liquid particles suspended in the air. PM with a diameter of 2.5 micrometers or less (PM2.5 - fine particles that can penetrate deep into lungs and bloodstream) and 10 micrometers or less (PM10 - coarse particles that can reach the lungs) are of particular concern for human health due to their ability to penetrate deep into the respiratory system.
- ✓ Ozone (O3): A gas that occurs both in the Earth's upper atmosphere (stratosphere), where it protects life from harmful ultraviolet radiation, and at ground level (troposphere), where it is a harmful air pollutant formed from reactions between other pollutants.
- ✓ Nitrogen oxides (NOx): a group of highly reactive gases including nitrogen dioxide and nitric oxide): A group of highly reactive gases containing nitrogen and oxygen, primarily formed during fuel combustion at high temperatures.
- ✓ Sulfur dioxide (SO2): A toxic gas produced by the burning of fossil fuels (coal and oil) and the smelting of mineral ore. It is a major contributor to acid rain and respiratory problems.

- ✓ Carbon monoxide (CO): A colorless, odorless, and tasteless gas produced by the incomplete burning of fuels.
- ✓ Volatile organic compounds (VOCs): Organic chemicals that have high vapor pressure at room temperature. Many VOCs are human-made chemicals used and produced in manufacturing processes, and they can react with nitrogen oxides in the presence of sunlight to form ground-level ozone.

The final dataset, a comprehensive compilation of these indicators, includes 217 rows (representing countries and years) and 19 columns (representing various pollutant metrics and associated data). Figure 1 provides an illustrative sample of data rows from the dataset.

	code	iso	country	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
0	876	WLF	Wallis and Futuna Islands	0.058569	0.056635	0.055689	0.054475	0.056712	0.052455	0.049982	0.050788	0.049598	0.050067	0.048616	0.048762	0.049999	0.048187	0.051362	0.050095
1	570	NIU	Niue	0.059838	0.058239	0.056967	0.055914	0.057780	0.053317	0.050798	0.051572	0.049778	0.051011	0.048137	0.048778	0.049933	0.048340	0.051866	0.050753
2	776	TON	Tonga	0.059927	0.058896	0.058301	0.056700	0.058044	0.053633	0.051075	0.052749	0.051293	0.051737	0.048889	0.049491	0.050606	0.049613	0.051790	0.050759
3	882	WSM	Samoa	0.060699	0.058907	0.058653	0.057342	0.059025	0.054557	0.051888	0.053584	0.051394	0.052414	0.050581	0.050653	0.052024	0.050672	0.052552	0.051334
4	184	COK	Cook Islands	0.059756	0.058554	0.057885	0.056303	0.058259	0.053418	0.050410	0.051857	0.049531	0.050904	0.047956	0.048062	0.049475	0.048105	0.052248	0.051419
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
212	156	CHN	China	0.647950	0.642482	0.639749	0.671731	0.696506	0.687931	0.695686	0.673202	0.689852	0.684705	0.703810	0.703645	0.704168	0.688977	0.689349	0.693457
213	50	BGD	Bangladesh	0.608520	0.583779	0.578247	0.644907	0.635473	0.631083	0.689169	0.617231	0.671047	0.669895	0.657703	0.707357	0.693170	0.698953	0.645853	0.735738
214	586	PAK	Pakistan	0.631339	0.643747	0.622211	0.623732	0.669380	0.644208	0.650113	0.710851	0.671764	0.677769	0.703974	0.707435	0.724600	0.778792	0.747174	0.757198
215	356	IND	India	0.632983	0.651811	0.620063	0.640686	0.652662	0.686162	0.657915	0.675458	0.701648	0.674637	0.680876	0.700919	0.728253	0.752007	0.731299	0.758780
216	524	NPL	Nepal	0.757440	0.779821	0.737915	0.783142	0.792879	0.818498	0.820017	0.813026	0.837248	0.823729	0.815767	0.826961	0.831658	0.912018	0.846573	0.899657

217 rows × 19 columns

Figure 1. The dataset

Among the pollutants, we focus specifically on **carbon monoxide (CO)** for deeper analysis. CO is a colorless, odorless, and tasteless gas produced by the incomplete burning of fuels such as wood, gasoline, and natural gas.

From a health perspective, CO poses serious risks due to its mechanism of action within the human body. It binds to **hemoglobin**—the protein molecule in red blood cells responsible for transporting oxygen from the lungs to the body's tissues and organs—much more strongly than oxygen itself. This preferential binding forms **carboxyhemoglobin (COHb)**, effectively preventing oxygen from reaching vital body tissues and organs. Prolonged or high exposure to CO can lead to acute symptoms such as dizziness, headache, nausea, and unconsciousness. Severe exposure can result in significant organ damage (particularly to the brain and heart) and even death. Furthermore, long-term exposure to lower levels of CO may increase the risk of heart disease and other chronic health conditions.

In the following analysis, we will explore the quantified measurements of carbon monoxide (CO) levels across various countries, utilizing visual representations for detailed insights into spatial and temporal trends.

3.1 Data analysis

This study was conducted using Google Colab, a cloud-based Python development environment that enables seamless execution of code and visualization of results.

The following Python libraries were used throughout the analysis:

- ✓ Pandas: for reading, cleaning, and manipulating tabular data.
- ✓ NumPy: for handling numerical operations and array structures.
- ✓ Plotly Express and Plotly Graph Objects: for creating interactive and dynamic visualizations.
- ✓ Matplotlib and Seaborn: for static plots and detailed visual analysis, including heatmaps and trend lines.

Table 1. Variable names, definitions, and roles in the machine learning model.

Variable Name (Abbreviation)	Variable Definition	Usage in the Study
Country	Country name. Each row represents a different country.	Included in the model as a categorical variable.
2003-2018	Annual air quality data for each year. Typically represents pollution concentration (e.g., PM2.5 or similar air quality metric).	Target variable to be predicted (dependent variable).

Table 2. The countries with the lowest CO values.

Country	Average CO	Average CO Percentage
Wallis and Futuna Islands	0.051999	5.199925
Niue	0.052689	5.268874
Cook Islands	0.052759	5.275900
Tonga	0.053344	5.334401
French Polynesia	0.053557	5.355697
Samoa	0.054142	5.414236
Falkland Islands	0.057541	5.754114
Solomon Islands	0.057541	5.764043
Nauru	0.058749	5.874852
New Caledonia	0.058862	5.886169

Table 3. The countries with the highest CO values.

Country	Average CO	Average CO Percentage
Nepal	0.818522	81.852184
Pakistan	0.685268	68.526799
India	0.684135	68.413492
China	0.682075	68.207500
North Korea	0.672977	67.297712
Bangladesh	0.654258	65.425785
South Korea	0.597307	59.730713
Rwanda	0.528443	52.844309
Laos	0.467295	46.729528
Bhutan	0.445680	44.568031

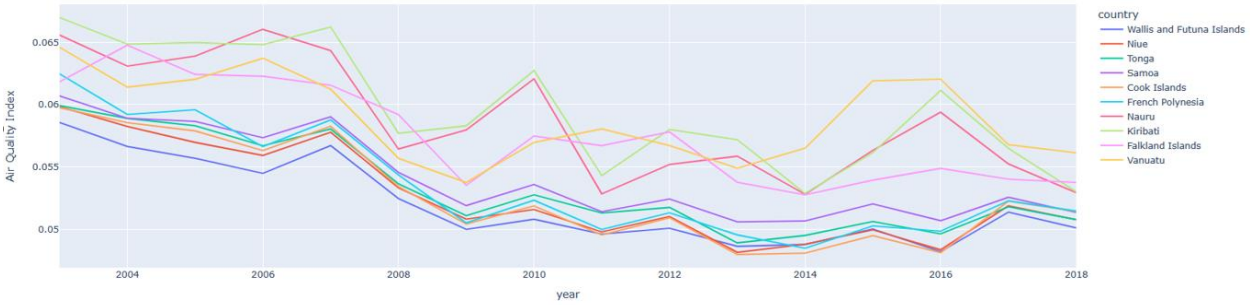

Figure 2. Change in CO values over the years (for countries with the lowest CO values)

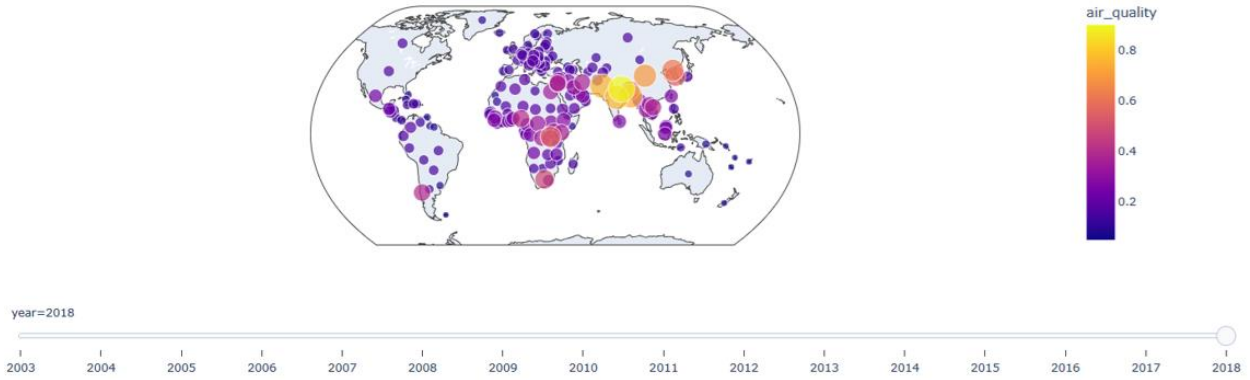

Figure 3. CO changes throughout the world over the years

In table 1, the variables corresponding to the years 2003 through 2018 represent annual air quality indicators and were treated as the target (dependent) variables in the analysis. The primary objective of the model was to predict or examine the temporal trends in air quality across these years. Before proceeding to the model development phase, the dataset was initially explored. Countries with the lowest and highest carbon monoxide (CO) values were identified, and their CO level trends over the years were analyzed. Table 2 presents the countries with the lowest CO values, while Table 3 highlights the countries with the highest CO values.

As seen in Figure 2, CO values in the countries with the lowest CO values have decreased significantly over the years. Figure 3 shows the temporal and spatial changes in CO2 levels worldwide between 2003 and 2018. Each circle on the map represents the CO2 concentration in a specific region. The color scale and circle size reflect air quality values. Purple and blue shades indicate low CO2 levels, while yellow and orange shades indicate high CO2 levels. The size of the circles indicates the magnitude of CO2 concentration observed in the relevant region. When examining the image over time, high CO2 levels are prominent in South Asia, Central Africa, and some industrial regions. Other regions experienced relatively low levels. This image visualizes the distribution of air pollution on a global scale and its annual trends, revealing the spatial dynamics of CO2 emissions.

3.2 Creating the model

Based on the comprehensive literature review, four machine learning algorithms were strategically selected for this study: Adaboost, XGBoost, Decision Tree, and Random Forest. The selection of these specific

algorithms was guided by their proven effectiveness in air quality prediction tasks as demonstrated in previous research.

XGBoost was chosen as a primary algorithm, as demonstrated in [27] and [23], who identified gradient boosting variants as consistently demonstrating superior performance across multiple air quality forecasting studies. We selected this algorithm as it showed exceptional results with correlation coefficients reaching high values in air quality applications, similar to the findings reported in these studies.

Random Forest was selected based on its proven effectiveness in air quality prediction, as shown in [16] and [10]. Lei et al. [16] demonstrated that ensemble methods, including Random Forest, show competitive performance in particulate matter prediction tasks, while Gupta et al. [10] achieved significant improvements in air quality classification accuracy. We chose this method following their approach, as it demonstrated strong generalization capabilities across different environmental conditions.

Adaboost was included in our model comparison, as used in the studies by [27] and [23], who emphasized the importance of comparing various boosting algorithms in air quality prediction. Although Ravindiran et al. [23] found AdaBoost to be less effective compared to other gradient boosting variants, we included it following their comparative methodology to ensure comprehensive performance evaluation.

Decision Tree was incorporated following the approach used by [5] and [1], who demonstrated the value of including traditional algorithms alongside ensemble methods in air quality studies. Like these studies, we included Decision Tree to understand the contribution of model complexity to prediction accuracy and to provide baseline comparison for ensemble approaches.

3.2.1 Adaboost model

Adaboost (Adaptive Boosting) is an ensemble learning algorithm that combines the outputs of multiple weak learners to create a strong predictive model [9]. The implementation follows the approach validated by [27] and [23] in their air quality forecasting studies. The algorithm sequentially trains weak learners, giving more emphasis to instances misclassified by previous learners. The final model is a weighted sum of weak learners, where each learner's weight is determined by its accuracy.

To create the AdaBoost model, equal weights were first assigned to all training examples. For a dataset containing N samples, each sample was initially given a weight value of 1/N. In the second step, a weak learner was trained, and the error was calculated by summing the weights of misclassified examples. The learner weight was calculated using:

$$at = \frac{1}{2} ln\,(1 - \frac{error_t}{error_t}) \quad (1)$$

Where $error_t$ is the weighted error rate of the weak learner at iteration t.

The weights of training instances were updated based on their performance using:

$$wi, t + 1 = wi, t \times exp(-at \times yi \times ht(xi)) \quad (2)$$

Where $w_{i,t+1}$ is the weight of instance i at iteration t+1, y_i is the true label, and $h_t(x_i)$ is the prediction of the weak learner (Hastie et al., 2009).

Performance metrics: MSE: 0.00018, R^2: 0.9879, RMSE: 0.0135, MAE: 0.0103

3.2.2 XGBoost model

XGBoost (Extreme Gradient Boosting) is an ensemble learning algorithm from the gradient boosting family [6]. This algorithm was selected based on its demonstrated superiority in multiple air quality studies, particularly those [27], who identified gradient boosting variants as consistently outperforming other approaches in

environmental prediction tasks. The model was created by first initializing parameters including learning rate (η), maximum tree depth, number of boosting rounds, and regularization parameters (γ and λ).

The loss function for regression was computed as:

$$L(yi, y^i) = \frac{1}{2}(yi - y^i)^2 \quad (3)$$

The predicted values were updated after adding the t-th tree:

$$y^i, t = y^i, t - 1 + \eta * ht(xi) \quad (4)$$

Regularization was implemented using:

$$\Omega f_k = \Upsilon T + \frac{1}{2}\lambda \sum_{J=1}^{T} w_j^2 \quad (5)$$

Where T is the number of leaves, w_j is the weight of leaf j, and γ, λ are regularization parameters.

The final prediction combines all weak learners:

$$H(x) = \sum_{t=1}^{T} \eta * h_{t(x)} \quad (6)$$

Performance metrics: MSE: 8.46×10^{-5}, R^2: 0.9944, RMSE: 0.0092, MAE: 0.0045

3.2.3 Random Forest model

Random Forest is an ensemble method that builds multiple decision trees and outputs the mode (classification) or mean (regression) of individual tree predictions [4]. The selection of this algorithm is supported by [16], who demonstrated competitive performance of ensemble methods in air quality prediction, and [10]. who showed significant improvements in air quality classification accuracy using Random Forest approaches.

For each tree, random feature subsets were selected at each split, and nodes were split based on the best feature among the random subsets. For classification, each tree votes using:

$$\hat{y} = argmaxi \sum_{J=1}^{N} (x) 1(hj(x) = i) \quad (7)$$

For regression, the final prediction averages all tree predictions:

$$\hat{y} = \frac{1}{N}\sum_{J=1}^{N} h_j(x) \quad (8)$$

Performance metrics: MSE: 8.49×10^{-5}, R^2: 0.9944, RMSE: 0.0092, MAE: 0.0040

3.2.4 Decision tree model

Decision trees recursively split data based on features to create a tree structure, using impurity measures such as Gini impurity or entropy as splitting criteria [22]. The inclusion of Decision Tree follows the comprehensive evaluation approach recommended by [5] and [1] who emphasized the importance of comparing individual algorithms with ensemble methods to understand the contribution of model complexity to predictive performance.

Gini impurity is calculated as:

$$I_G(t) = 1 - \sum_{i=1}^{c} p(i|t)^2 \quad (9)$$

Entropy is calculated as:

$$I \in (t) = -\sum_{j=1}^{c} P(t) * (p(i|t)) \quad (10)$$

For classification, the majority class in a leaf node becomes the prediction:

$$\hat{y} = argmaxi \sum_{j=1}^{N} (x) 1(hj(x) = i) \quad (11)$$

For regression, the mean value of target values in a leaf node is used:

$$\hat{y} = \frac{1}{N}\sum_{J=1}^{N} h_j(x) \quad (12)$$

Performance metrics: MSE: 0.0012, R^2: 0.9184, RMSE: 0.0351, MAE: 0.0098

3.2.5 Predictions

When four different models created with four different methods were compared with each other in terms of performance metrics, it was concluded that the model with the highest performance was XGBoost. Based on this situation, predictions for the coming years were made using the model created with the XGBoost method. Based on the predictions for countries, the highest CO level changes and the lowest CO level changes were calculated for the period 2019-2028. Then, the countries that are likely to experience the worst changes were investigated. Table 3 shows the countries with the highest CO level changes; Table 4 shows the countries with the lowest CO level changes.

According to the prediction results provided in Table 3, CO levels will increase in Bhutan, South Korea, North Korea, and China. Bhutan and North Korea will have the biggest absolute increases. On the other hand, South Africa, Bangladesh, Pakistan, India, and Nepal will decrease their CO levels. India, Pakistan, and Nepal will experience significant decreases in CO levels compared to other countries. These countries may have improvements in CO levels; however, they will remain high. Figure 4 provides a visual representation of the predictions.

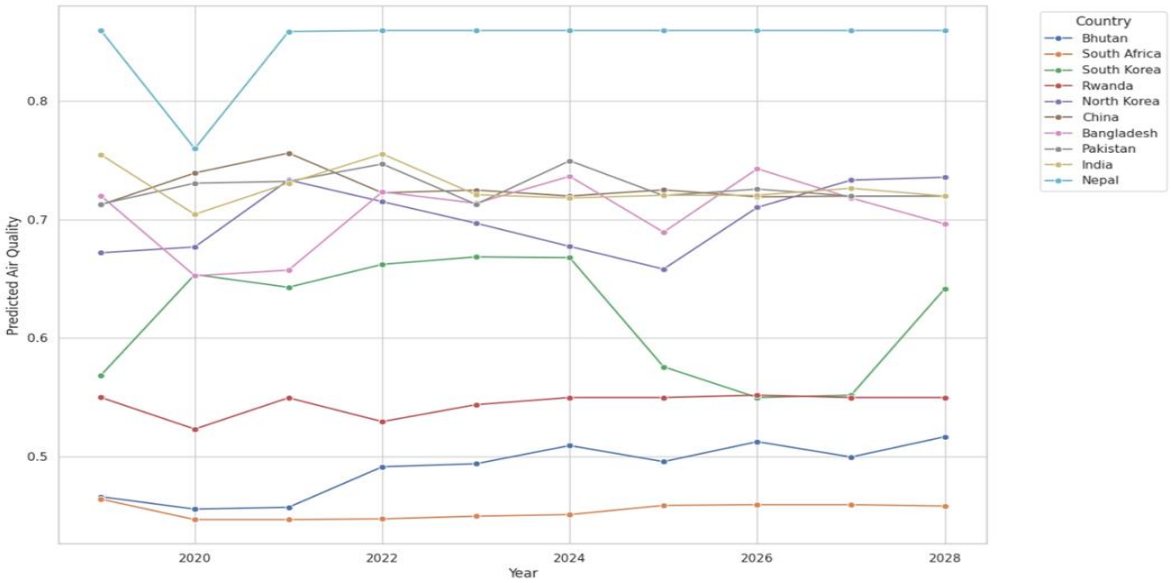

Figure 4. Predicted air quality changes (2019-2028) for countries with highest CO levels

According to the prediction results provided in Table 4, Wallis and Futuna Islands, Niue, Tonga, and Samoa will have minimal changes in CO levels. Wallis and Futuna Islands will have the smallest increase. Over time, these regions will maintain relatively stable CO levels. French Polynesia and Nauru will have slight increases, but the changes are relatively small. This stability may indicate consistent air quality management or low pollution levels in these regions. Figure 5 provides a visual representation of the predictions.

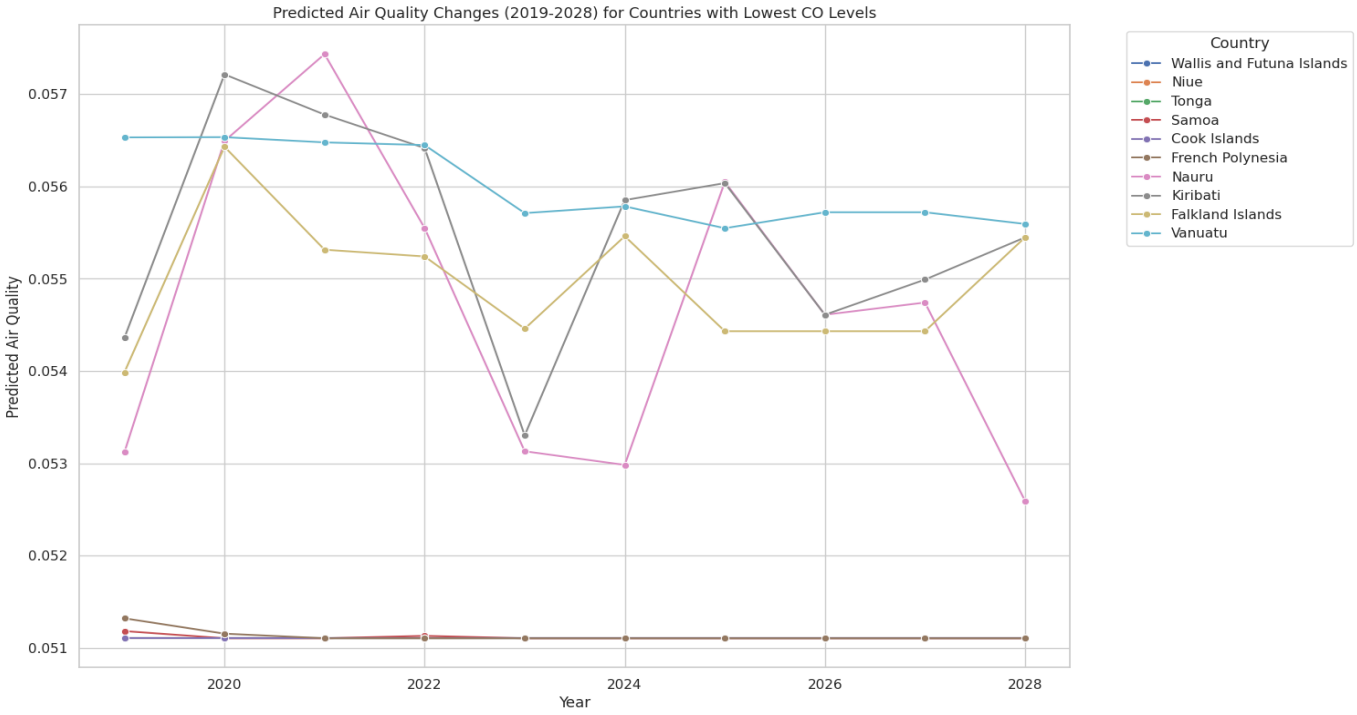

Figure 5. Predicted air quality changes (2019-2028) for countries with lowest CO levels

Overall, the analysis highlights significant variations in CO levels changes across countries. Bhutan and North Korea, with the highest increases, and countries such as India, Pakistan, and Nepal, with noticeable decreases, reflect diverse trends in air quality.

3.2.6 Environmental justice implications of predicted air quality trends

The projected CO-level increases in countries such as Laos (+0.183878 from 2019 to 2028), Indonesia (+0.116214), and North Korea (+0.114642) clearly demonstrate a pattern of environmental injustice. Specifically, these nations contribute comparatively small shares of global emissions yet are forecast to bear disproportionately high burdens of deteriorating air quality. For example, Laos—although historically a minor industrial emitter—faces the largest absolute CO increase, highlighting an inequitable distribution of environmental risk.

Furthermore, countries with already elevated baseline CO levels (e.g., North Korea: 0.620899 in 2018) are predicted to deteriorate further, thereby compounding existing disadvantages. This compounding effect widens the justice gap, because populations in these regions generally lack the economic capacity or healthcare infrastructure needed to adapt or mitigate such outcomes.

An inverse relationship emerges between adaptive capacity and severity of impact. Nations with limited financial and institutional resources (e.g., Laos, Bhutan) are least able to invest in emission controls or public health interventions, highlighting systemic environmental injustice at a global scale.

Finally, these predicted patterns carry major health equity implications:

- ✓ Indonesia's large population (≈270 million) facing a +0.116214 CO increase portends millions of individuals at heightened risk of respiratory illness.
- ✓ Laos and Bhutan, which already struggle with limited healthcare infrastructure, will see a compounding burden of

environmental degradation coinciding with insufficient healthcare response capacity.

- ✓ Intergenerational equity is also at stake, as future cohorts in severely affected regions will inherit disproportionately degraded air quality, thus perpetuating the cycle of inequality.

Together, these findings underscore that environmental justice is not merely a local or regional issue, but a global one. The quantitative predictions must therefore inform targeted international policies and resource allocation, particularly for the most vulnerable populations.

4 Results

We will first explain the performance metrics of each machine learning model used in predicting air quality, followed by a discussion of the key findings, insights, and spatial patterns uncovered by these models. The machine learning models demonstrated high predictive accuracy, as indicated by metrics such as mean squared error (MSE), R-squared (R^2), root mean squared error (RMSE), and mean absolute error (MAE). Table 5 presents the performance metrics for each model.

The comprehensive evaluation of machine learning algorithms in air quality prediction reveals significant variations in performance across different studies and methodologies. A detailed comparison between the current study's findings and previous research demonstrates both alignments and notable improvements in predictive accuracy.

The comparative analysis reveals several important insights regarding the performance of machine learning algorithms in air quality prediction. The current study's Random Forest and XGBoost models demonstrate exceptional performance with R^2 values exceeding 0.994, which surpasses most previous studies. This represents a significant improvement over MLR models with R^2 = 0.88–0.90 [16] and LSTM models with R^2 = 0.9137 [17]. The achievement is particularly notable when compared to a stacked regressor approach which achieved an R^2 of 0.973 with considerably higher RMSE (7.568) and MAE (4.596) values [1].

The superior performance of Random Forest and XGBoost in the current study aligns with findings from multiple previous studies that highlighted the effectiveness of ensemble methods (techniques that combine multiple algorithms). LightGBM and CatBoost have been identified as top performers among boosting algorithms with R^2 up to 0.9998 [23, 27]. However, the current study's XGBoost implementation achieves comparable R^2 performance (0.9944) with significantly lower error metrics [27].

The relatively weaker performance of Adaboost in the current study (R^2 = 0.9879) is consistent with findings from previous works that also identified Adaboost as the least effective among boosting algorithms [23, 27]. This consistency across different datasets and geographic regions suggests inherent limitations of Adaboost for air quality prediction tasks.

The current study achieves remarkably low error metrics, with RMSE values ranging from 0.009 to 0.035, which represent substantial improvements over previous studies. For comparison, RMSE values between 4.26 and 7.65 were reported [16], while RMSE values ranging from 0.1403 to 0.5674 were observed for Random Forest implementations across different cities [10].

Table 3. Countries with the highest CO level changes (Actual value: 2018, Predictions: 2019-2028).

County	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028
Bhutan	0.466100	0.466150	0.455674	0.457219	0.491414	0.493955	0.509318	0.495733	0.512580	0.499620	0.516732
South Africa	0.488000	0.464094	0.446874	0.446874	0.447484	0.449714	0.451116	0.458802	0.459449	0.459449	0.458312
South Korea	0.535970	0.568631	0.653514	0.642788	0.662117	0.668427	0.667762	0.575752	0.549816	0.551783	0.641697
Rwanda	0.549030	0.549843	0.523209	0.549581	0.529495	0.543791	0.549816	0.549816	0.551783	0.549816	0.549816
North Korea	0.620899	0.671910	0.676762	0.733278	0.715022	0.696841	0.677234	0.658135	0.710186	0.733148	0.735542
China	0.693456	0.712349	0.739179	0.755998	0.722484	0.724756	0.719769	0.724952	0.719043	0.719696	0.719696
Bangladesh	0.735738	0.735738	0.652480	0.657237	0.723244	0.723244	0.736357	0.689264	0.689264	0.718079	0.696166
Pakistan	0.757198	0.712949	0.712949	0.732116	0.746784	0.712764	0.749499	0.720065	0.725605	0.719696	0.719696
India	0.758780	0.754564	0.704334	0.730682	0.755280	0.720909	0.718122	0.720422	0.720422	0.726331	0.719696
Nepal	0.899657	0.899657	0.759672	0.858449	0.859298	0.859298	0.859298	0.859298	0.859298	0.859298	0.859298

Table 4. Countries with the lowest CO level changes (Actual value: 2018, Predictions: 2019-2028).

County	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028
Wallis Futuna Island	0.050094	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088
Niue	0.050752	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088
Tonga	0.050758	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088
Samoa	0.051333	0.051166	0.051088	0.051114	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088
Cook Islands	0.051418	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088
French Polynesia	0.051447	0.051326	0.051111	0.051114	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088	0.051088
Nauru	0.052925	0.053119	0.056623	0.057439	0.055494	0.053366	0.053238	0.056024	0.054412	0.054310	0.055379
Kiribati	0.053002	0.054418	0.057082	0.056758	0.056252	0.053561	0.055769	0.056024	0.054412	0.054501	0.055379
Falkland Islands	0.053748	0.053928	0.056514	0.055230	0.055167	0.054133	0.055379	0.054109	0.054109	0.054109	0.055379
Vanuatu	0.056130	0.056655	0.056553	0.056581	0.056429	0.055607	0.055676	0.055472	0.055472	0.055472	0.055494

Table 5. Countries with the highest CO level changes (Actual Values: 2019 - 2024).

Country	2019	2020	2021	2022	2023	2024
Bhutan	50.3%	39.9%	58.3%	73.5%	29.9%	87.0%
South Africa	38.8%	39.7%	30.2%	25.3%	29.0%	25.9%
South Korea	50.2%	52.4%	25.8%	21.5%	36.5%	31.9%
Rwanda	62.7%	58.2%	38.0%	31.7%	36.5%	32.7%
North Korea	46.7%	49.4%	38.5%	32.2%	37.0%	32.8%
China	62.7%	58.2%	57.9%	47.0%	38.8%	46.2%
Bangladesh	84.3%	70.9%	75.6%	73.5%	83.6%	87.0%
Pakistan	88.5%	71.9%	86.9%	79.2%	92.3%	80.3%
India	88.5%	80.8%	75.5%	58.4%	68.1%	55.2%
Nepal	95.7%	76.3%	57.9%	47.0%	53.6%	46.2%

Table 6. Countries with no publicly available CO level data due to discontinued publication by the provider (2019–2024).

Country	2019	2020	2021	2022	2023	2024
Wallis Futuna Island	N/A	N/A	N/A	N/A	N/A	N/A
Niue	N/A	N/A	N/A	N/A	N/A	N/A
Tonga	N/A	N/A	N/A	N/A	N/A	N/A
Samoa	N/A	N/A	N/A	N/A	N/A	N/A
Cook Islands	N/A	N/A	N/A	N/A	N/A	N/A
French Polynesia	N/A	N/A	N/A	N/A	N/A	N/A
Nauru	N/A	N/A	N/A	N/A	N/A	N/A
Kiribati	N/A	N/A	N/A	N/A	N/A	N/A
Falkland Islands	N/A	N/A	N/A	N/A	N/A	N/A
Vanuatu	N/A	N/A	N/A	N/A	N/A	N/A

Table 7. Comparison of the current study models.

Method	Mean Squared Error (MSE)	R-Squared (R2)	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)
Decision Tree	0.001200	0.918400	0.035100	0.000900
Random Forest	8.49e-05	0.004383	0.009214	0.003962
XGBoost	8.46e-05	0.994403	0.009198	0.004468
Adaboost	0.000183	0.987872	0.013540	0.010255

Table 8. Comparison of the models with previous studies.

Method	Mean Squared Error (MSE)	R-Squared (R2)	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)
MLR	-	0.88-0.90	4.26-7.65	2.84-4.73
CatBoost	0.580000	0.999800	0.760000	0.60000
Adaboost	-	0.975300	-	-
Random Forest	-	-	0.567400	-
LSTM	-	0.913700	-	-
Stacked Regressor	-	0.973000	7.568000	4.596000
Time Series	166.358000	-	-	-

Table 9. The worst changes (Actual value: 2018, Prediction: 2028)

Country	Change	2018	2028
Laos	0.183878	0.339147	0.523025
Indonesia	0.116214	0.266231	0.382445
North Korea	0.114642	0.620899	0.735542
South Korea	0.105727	0.535970	0.641697
Bhutan	0.050546	0.466186	0.516732
China	0.026240	0.693457	0.719696
Lebanon	0.024324	0.308343	0.332667
Sierra Leone	0.016519	0.268831	0.285351
Singapore	0.016280	0.175256	0.191537
Egypt	0.013091	0.342315	0.355407

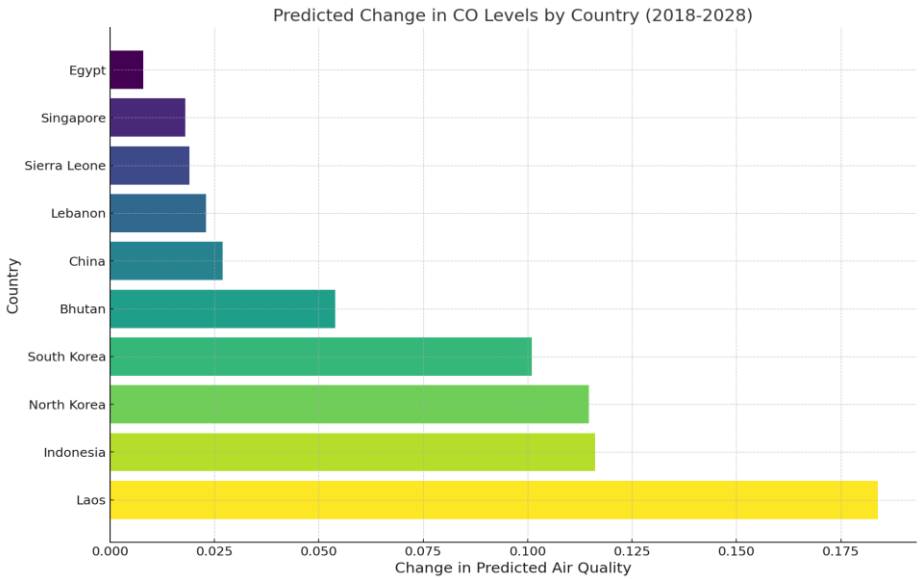

Figure 6. Countries with the worst changes in air quality (Actual value: 2018, Prediction: 2028)

In Table 6, no actual data are publicly available for these countries because the data provider has discontinued publishing such information. Consequently, these entries are indicated as N/A. The predictions in this study cover the period from 2019 to 2028 and are based on actual data from 2018. The provider has released real data only up to 2024, and the lack of subsequent data is due to the provider's decision to cease publication.

It appears that XGBoost is indeed performing well among the considered machine learning models for air quality prediction tasks. Its combination of low mean squared error, high R-squared, low root mean squared and mean absolute error values indicated its effectiveness in capturing and predicting air quality variations. The model was trained using the following hyperparameters (settings that control the learning process):

- ✓ n_estimator: 300
- ✓ learning_rate = 0.05
- ✓ max_depth = 7
- ✓ subsample = 0.8
- ✓ colsample_bytree = 0.9
- ✓ reg_alpha = 0.1
- ✓ reg_lambda = 1.5

These values were chosen based on prior testing to balance model complexity and generalization performance.

Therefore, we used the XGBoost model to carry out the predictions. We created table 3 and 4 by investigating the countries with the lowest and highest changes.

The analysis reveals that Laos, Indonesia, and North Korea will experience the most crucial changes in air quality, with changes of 0.183878, 0.116214, and 0.114642, respectively. These countries will have notable decreases in their air quality Actual value: 2018, Prediction: 2028. Figure 6 presents a visual representation of the air quality changes of these countries.

Our study serves as a springboard for future research endeavors that delve deeper into the complexities of air quality dynamics. While our current models incorporate a comprehensive set of variables, the integration of specific factors could further enrich our predictive power, which may include detailed land-use patterns, industrial emissions, or localized weather phenomena. Tailoring machine learning models to forecast the concentration of individual pollutants enables a targeted approach to environmental management and public health interventions. This focused analysis could uncover distinct trends, sources, and health implications associated with specific pollutants, contributing to a more nuanced understanding of air quality challenges.

The analysis reveals that developing countries, particularly in Southeast Asia, face the most severe air quality deterioration. Laos and Indonesia, despite contributing less to global emissions historically, are projected to experience the largest air quality degradation. This exemplifies environmental injustice where nations with limited industrial development bear disproportionate environmental burdens.

Countries already experiencing poor air quality (North Korea: 0.621, South Korea: 0.536) are predicted to deteriorate further, creating a compounding effect that widens the environmental justice gap. This pattern suggests that existing environmental disadvantages become self-perpetuating without target intervention.

The results demonstrate an inverse relationship between adaptive capacity and projected impact severity. Nations with limited economic resources (Laos, Bhutan) face proportionally greater challenges in addressing air quality deterioration, highlighting systemic environmental injustice at the global scale.

The concentration of severely affected countries in Southeast Asia indicated regional environmental justice concerns, where

geographical proximity to pollution sources, prevailing wind patterns, and shared atmospheric conditions create collective disadvantages.

The projected air quality deterioration patterns directly translate to significant health equity concerns across multiple dimensions. Indonesia's large population of over 270 million facing a 0.116 deterioration coefficient represents massive public health implications, as millions of individuals will experience worsened respiratory conditions and related health outcomes. Countries like Laos with limited healthcare infrastructure face dual burdens of worsening air quality and inadequate health response capacity, creating a compounding effect where environmental degradation intersects with healthcare system limitations. Furthermore, current deterioration trends raise serious intergenerational justice concerns. as future generations in the most affected regions will inherit disproportionately compromised conditions that will affect their health outcomes throughout their lifespans.

The above results clearly demonstrate that predicted changes in CO concentrations are tied directly to environmental justice concerns. In the subsequent discussion, we interpret these implications for policy and future research.

5 Discussion and conclusion

The countries with the least change in CO levels and the best air quality are all islands. These islands are in the Pacific, Atlantic and Pacific Oceans. The countries with the worst air quality and the highest possible increases in CO levels are shown in Figure 7.

In conclusion, our analysis confirms that environmental injustice remains a global challenge: nations with minimal historic emissions (e.g., Laos, Indonesia) are forecast to experience the largest CO increases, while countries already suffering poor air quality (e.g., North Korea, South Korea) face further deterioration. The projected inverse relationship between adaptive capacity and impact severity (most vividly seen in resource-limited nations like Laos and Bhutan) highlights systemic inequities that require urgent attention.

5.1 Real-world impacts on marginalized communities

While our analysis highlights projected country-level shifts in CO concentrations, understanding how these changes translate into localized, on-the-ground injustices requires examining specific communities. In many urban and rural settings, low-income and minority groups experience a "double burden": they face higher exposure to pollution and simultaneously have less capacity to mitigate health risks. Below, we present two illustrative case studies—and summarize related literature—to show how poor air quality disproportionately harms marginalized populations.

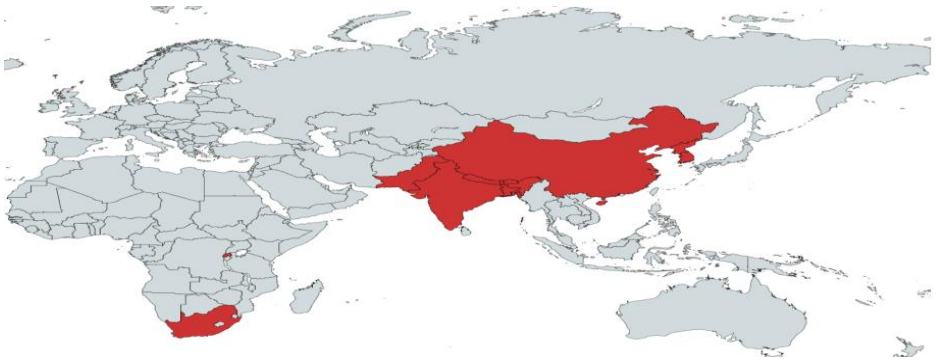

Figure 7. Countries with the worst air quality

5.2 Urban low-income neighborhoods

In major megacities of South and Southeast Asia—such as Dhaka (Bangladesh) and Jakarta (Indonesia)—informal settlements and low-income districts are often situated adjacent to industrial corridors, waste-burning areas, or heavily trafficked roads [24, 7]. For example, daily $PM_{2.5}$ and CO measurements in Dhaka's poorest wards regularly exceed World Health Organization limits by two- to threefold, whereas wealthier districts often remain within or near acceptable ranges. Residents of these communities tend to work in unregulated factories, street vending, or informal transportation, spending extended periods outdoors without access to effective indoor air filtration. Consequently, pediatric asthma rates in Dhaka's low-income neighborhoods are reported to be roughly 15 % higher than city-wide averages, and chronic obstructive pulmonary disease (COPD) incidence is nearly double in Jakarta's informal resettlement areas [23]. These disparities illustrate how air pollution both reflects and perpetuates socio-economic inequities—an environmental injustice seen at the neighborhood level.

5.3 Rural and indigenous communities

Outside urban centers, rural and indigenous populations also endure disproportionate air-quality burdens. In inland Chinese provinces (e.g., Guizhou, Shanxi), many households rely on coal-fired stoves for cooking and heating, producing indoor CO concentrations as high as 9 ppm—more than twice recommended limits—according to Li et al. (2018). Women and children, who spend more time near these stoves, exhibit elevated blood carboxyhemoglobin (COHb) levels and a twofold increase in pediatric pneumonia. Similarly, indigenous communities in northern Mexico endure high $PM_{2.5}$ and CO from seasonal agricultural burning; hospitalization rates for acute respiratory infections in these regions are 1.8 times higher than in nearby non-indigenous towns (Martínez-Galán et al., 2020). These case studies show that, even if national average CO levels seem moderate, specific rural and indigenous populations remain at acute risk—a manifestation of environmental injustice that national-level models alone may not capture.

5.4 Synthesis and policy implications

Together, these community-level examples demonstrate that elevated pollutant levels do not distribute evenly across society. Low-income urban residents and rural households reliant on biomass or coal face higher exposures due to proximity to emission sources or lack of clean-fuel alternatives—and they often lack access to preventive healthcare, public information on air quality, or the financial means to adopt cleaner technologies. International studies indicate that air pollution–related mortality and morbidity are 1.5–2 times higher among vulnerable groups—children, the elderly, and those without robust healthcare coverage [26, 21]. As a result, even if a country's overall CO concentrations decline, the benefits may accrue unevenly, leaving marginalized communities behind.

To address these injustices, targeted interventions—such as subsidized clean-cooking stoves, stricter zoning laws near industrial sites, and community-based air monitoring—are urgently needed. Future research should combine our machine-learning forecasts with high-resolution, community-level data to identify specific neighborhoods or villages at greatest risk, thereby guiding equitable policy measures.

5.5 Underlying drivers of predicted CO changes

While our analysis forecasts country-level CO trends, it is crucial to understand the real-world factors that drive these changes. Below, we discuss key socio-economic, industrial, and policy-related determinants for two illustrative cases—Bhutan and North Korea—and then summarize broader considerations that affect CO trajectories across all countries.

Bhutan: Bhutan's development model has historically emphasized environmental conservation. Over 70% of the nation's electricity is generated by run-of-river hydropower, which keeps domestic CO emissions relatively low [3]. Nevertheless, our model predicts a modest increase in Bhutan's CO levels (approximately +0.012 ppm from 2019 to 2028). Several factors contribute:

- ✓ **Rising vehicle ownership:** As household incomes gradually rise, private vehicle registrations have grown at an average rate of ~5% per year [26]. Increased passenger cars and motorcycles lead to higher on-road fossil fuel combustion, offsetting some gains from hydropower.
- ✓ **Construction sector expansion:** Ongoing hydropower projects and new infrastructure in urban and peri-urban areas generate additional diesel generator usage and construction-equipment emissions. For example, large dam projects in western Bhutan have increased diesel consumption at site camps by up to 20% during peak construction months [8].
- ✓ **Transboundary pollution:** Seasonal winds carry pollutants from northern India's coal-fired power plants into southern Bhutan. Up to 15% of ambient CO measured in Thimphu during winter months originates from cross-border sources [21]. Although Bhutan enforces strict domestic emission controls, it cannot fully eliminate these imported pollutants.
- ✓ **Emerging policy responses:** To counterbalance these pressures, Bhutan introduced a nationwide ban on single-use plastic bags in 2018 and has begun subsidizing electric motorcycles in Thimphu [19]. Such policies may moderate future CO increases, but their full impact will depend on scale and enforcement.

In summary, Bhutan's projected CO uptick reflects a combination of increased vehicular emissions and construction activities—despite a hydropower-dominated grid—and the unavoidable influence of neighboring India's coal-driven pollution.

North Korea: North Korea is forecast to experience a substantial rise in CO levels (+0.1146 ppm from 2019 to 2028). Several interrelated factors explain this trend:

- ✓ **Outdated coal-fired infrastructure:** North Korea relies heavily on aging coal plants with minimal emission controls. Satellite-derived nighttime-light data and industrial output proxies show that coal consumption surged in 2018–2019 to compensate for power shortages, leading to elevated CO emissions [15].
- ✓ **Fuel shortages and biomass burning:** Economic sanctions and chronic fuel scarcity force some households to burn wood and charcoal in inefficient, unventilated stoves. Indoor CO concentrations in rural homes often exceed safe limits during winter heating, contributing to elevated ambient CO when household emissions mix with outdoors [14].
- ✓ **Small-scale industrial activities:** Defector interviews and limited field studies indicate that brick kilns and small metal-smelting workshops continue operating on low-grade coal or scrap during off-peak agricultural seasons. These informal industrial emissions cause seasonal CO spikes, particularly around regional manufacturing clusters [7].
- ✓ **Urban transportation emissions:** Pyongyang's public bus and trolleybus fleets remain largely diesel-powered and lack regular maintenance or emission testing. As ridership rebounded after 2015, on-road CO emissions increased due to the absence of robust vehicle inspection programs [20].
- ✓ **Weak regulatory framework:** Although North Korea enacted a "Clean Air Law" in the 1980s, enforcement has been virtually nonexistent for decades due to limited government capacity and resource

constraints. Consequently, there is no systematic monitoring of industrial or vehicular emissions, and no incentives to adopt cleaner fuel technologies.

Together, these factors—obsolete coal plants, household biomass burning, unregulated small industries, rising urban transport emissions, and a lack of effective air quality policy—explain why our model forecasts a significant CO increase in North Korea.

Beyond the country-specific cases above, several broader drivers influence CO trends globally:

- ✓ **Economic growth trajectories:** Gross domestic product (GDP) per capita strongly correlates with energy demand. Rapidly expanding economies often see sharp CO increases unless they simultaneously invest in cleaner energy [13]. Countries experiencing double-digit GDP growth—such as Vietnam and Myanmar—have shown corresponding upticks in domestic CO emissions.
- ✓ **Fuel mix transitions:** A shift from coal and biomass to cleaner fuels (e.g., liquefied natural gas, LPG, renewables) typically reduces CO concentration. For instance, Indonesia's partial conversion of power plants to natural gas in 2019 led to a 5% drop in CO emissions that same year [18]. Conversely, continued reliance on coal in countries like Pakistan and South Africa drives CO higher.
- ✓ **National air quality policies:** Countries with active air monitoring networks and strict vehicle-emission regulations (e.g., South Korea's Emissions Trading Scheme initiated in 2015) frequently achieve more rapid CO reductions than statistical trend models predict. By contrast, nations lacking regulatory frameworks or enforcement—such as Myanmar and Haiti—often experience unmitigated CO increases.
- ✓ **Urbanization and demographics:** Urban migration can raise CO through traffic, poor transit, and informal industries, while planned growth with clean transit and energy can offset impacts.
- ✓ **Transboundary and regional emission patterns:** Prevailing wind patterns and geographic proximity to large emitters can import CO pollution. For example, Nepal and Bangladesh receive wintertime haze from northern India's coal districts, temporarily elevating ambient CO despite domestic emission controls [24].

By integrating socio-economic, industrial, and policy factors, we explain why countries like Bhutan and North Korea show the CO trends in Section 3. Beyond advancing methods, policymakers must direct resources, technical aid, and interventions to vulnerable regions. Framing ML forecasts within environmental justice can guide fairer air quality policies, ensuring cleaner air and a more equitable distribution of environmental risks for future generations.

5.6 Policy integration and ML-enabled governance

Our machine-learning forecasts not only identify which countries and regions will face the greatest CO increases, but also offer a foundation for more proactive, data-driven air-quality management. Below, we outline how policymakers could utilize these predictions within environmental governance frameworks and suggest concrete interventions to mitigate adverse outcomes.

5.7 Embedding ML forecasts into regulatory planning:

- ✓ **Early warning and alerts:** Linking CO forecasts with real-time sensors (e.g., IoT monitors) enables dynamic thresholds that trigger alerts when levels exceed safe limits. In Bhutan, a dashboard could warn health authorities and schools when predicted CO tops 1.0 ppm, prompting measures like mask distribution or temporary closures. National regulations could mandate ML-powered early-warning dashboards tied to public health actions.

- ✓ **Dynamic emissions permitting:** Traditional emission permits often use static caps. Machine-learning projections enable "adaptive permitting," where allowable emissions fluctuate based on forecasted background CO levels. For example, in North Korea's industrial clusters, regulators—once ML forecasts are approved—could temporarily lower permitted CO emissions for brick kilns and small foundries during periods when ambient CO is predicted to spike, thereby preventing compounding pollution episodes. This approach can be enshrined in environmental statutes as a condition that any facility's permit must factor in ML-based background estimates. Noncompliance (i.e., failing to reduce actual emissions when forecasts exceed warning thresholds) could trigger fines or temporary shutdown orders.
- ✓ **Targeted resource allocation:** Our country-specific predictions can guide limited public-health budgets toward communities that are most at risk in the coming years. For example, for Lao PDR—projected to have the highest CO increase—the Ministry of Health can pre-position mobile clinics and stockpile inhalers and oxygen concentrators in the provinces forecast to exceed 1.2 ppm consistently. The government may establish a "Health-Environment Fund" that automatically releases grants or equipment to provinces whose ML-driven CO projections surpass defined thresholds. This rule could be written into fiscal policy guidelines, ensuring transparency.

5.8 Machine learning in compliance and enforcement:

- ✓ **Satellite and remote-sensing integration:** Merging ground-based CO forecasts with satellite data (e.g., TROPOMI) can reveal unreported emission hotspots. For instance, India–Bhutan border officials could flag areas where satellite CO exceeds ML forecasts by over 10%, triggering inspections. Annual audits could compare forecasts with satellite anomalies to target regulatory action.
- ✓ **Predictive maintenance for transport fleets:** Municipalities can use ML models trained on vehicle CO data to predict which buses or trucks will exceed limits before the next service. In Pyongyang, CO monitors on sample buses could flag units likely to exceed 1.5 ppm, prioritizing maintenance and meeting quarterly "ML-guided inspection" requirements to cut fleet emissions.

Acknowledgment

This project was carried out as an undergraduate graduation project in the Management Information Systems program of Işık University, Department of Information Technologies. The implementation of the project was carried out by Gorkem Demircan and was supervised by Dr. Gülsüm Çiğdem Çavdaroğlu.

Conflict of Interest

The authors of this article declare that there is no conflict of interest.

References

[1] S. A. Aram, E. A. Nketiah, B. M. Saalidong, H. Wang, A. R. Afitiri, A. B. Akoto, and P. O. Lartey, "Machine learning-based prediction of air quality index and air quality grade: a comparative analysis," *International Journal of Environmental Science and Technology*, vol. 21, no. 2, pp. 1345–1360, 2024.

[2] P. Bhalgat, S. Pitale, and S. Bhoite, "Air quality prediction using machine learning algorithms," *International Journal of Computer Applications Technology and Research*, vol. 8, no. 9, pp. 367–370, 2019.

[3] Bhutan Electricity Authority, *Bhutan energy statistics 2020*. Thimphu, Bhutan: Bhutan Electricity Authority, 2020.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] M. Castelli, F. M. Clemente, A. Popovič, S. Silva, and L. Vanneschi, "A machine learning approach to predict air quality in California," *Complexity*, 2020.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[7] Y. Choe and H. Kim, "Emissions from small-scale industries in North Korea: A field study," *East Asian Industrial Review*, vol. 3, no. 1, pp. 12–27, 2020.

[8] K. Dorji, P. Wangchuk, and T. Tshering, "Impact of hydropower construction on air quality in western Bhutan," *Journal of Himalayan Environmental Studies*, vol. 5, no. 2, pp. 45–58, 2019.

[9] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[10] N. S. Gupta, Y. Mohta, K. Heda, R. Armaan, B. Valarmathi, and G. Arulkumaran, "Prediction of air quality index using machine learning techniques: a comparative analysis," *Journal of Environmental and Public Health*, vol. 2023, no. 1, p. 4916267, 2023.

[11] M. Hardini, R. A. Sunarjo, M. Asfi, M. H. R. Chakim, and Y. P. A. Sanjaya, "Predicting air quality index using ensemble machine learning," *ADI Journal on Recent Innovation*, vol. 5, no. 1Sp, pp. 78–86, 2023.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*, 2nd ed. Springer, 2009.

[13] International Energy Agency, *World energy outlook 2021*. Paris, France: IEA, 2021.

[14] D. S. Kim and J. Y. Park, "Household energy use and indoor air quality in rural DPRK," *Journal of Asian Public Health*, vol. 10, no. 4, pp. 210–222, 2019.

[15] S. H. Lee, M. J. Kang, and H. J. Park, "Satellite-based assessment of coal consumption in North Korea," *International Journal of Remote Sensing*, vol. 39, no. 14, pp. 4768–4782, 2018.

[16] T. M. Lei, S. W. Siu, J. Monjardino, L. Mendes, and F. Ferreira, "Using machine learning methods to forecast air quality: A case study in Macao," *Atmosphere*, vol. 13, no. 9, p. 1412, 2022.

[17] Q. Liu, B. Cui, and Z. Liu, "Air quality class prediction using machine learning methods based on monitoring data and secondary modeling," *Atmosphere*, vol. 15, no. 5, p. 553, 2024.

[18] Ministry of Energy and Mineral Resources, Indonesia, *Annual energy report 2020: Fuel consumption and emissions*. Jakarta, Indonesia, 2020.

[19] Ministry of Environment, Bhutan, *Policy on single-use plastics and electric vehicle promotion*. Thimphu, Bhutan, 2019.

[20] J. H. Park, M. K. Lee, and S. Y. Choi, "Evaluating public transportation emissions in Pyongyang, vol. 7, no. 2, pp. 101–113, 2020.

[21] T. Phuntsho and Y. Tashi, "Transboundary air pollution from India to Bhutan: An analysis," *Environmental Science and Policy Journal*, vol. 12, no. 1, pp. 23–34, 2020.

[22] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[23] G. Ravindiran, G. Hayder, K. Kanagarathinam, A. Alagumalai, and C. Sonne, "Air quality prediction by machine learning models: A

predictive study on the Indian coastal city of Visakhapatnam," *Chemosphere*, vol. 338, p. 139518, 2023.

[24] United Nations Environment Programme, *Air pollution in South Asia: Transboundary haze report*. Nairobi, Kenya, 2020.

[25] S. Wang, J. McGibbon, and Y. Zhang, "Predicting high-resolution air quality using machine learning: Integration of large eddy simulation and urban morphology data," *Environmental Pollution*, vol. 344, p. 123371, 2024.

[26] World Bank, *World Development Indicators: Transport statistics*. Washington, DC: World Bank, 2021.

[27] Y. Zhang, Y. Wang, M. Gao, Q. Ma, J. Zhao, R. Zhang, and L. Huang, "A predictive data features exploration-based air quality prediction approach," *IEEE Access*, vol. 7, pp. 30732–30743, 2019.