Data Augmentation Techniques for Improved $PM_{2.5}$ Forecasting Using Transformer Architectures

Phoebe Pan

05/27/2025

Yilmaz - Period 3

Table of Contents

Abstract

Exposure to fine particulate matter with diameter less than 2.5 μm (PM2.5) significantly increases an individual's risk of cardiovascular and respiratory disease. As climate change progresses, extreme events, including wildfires, are projected to increase, exacerbating air pollution. However, models often struggle to capture extreme pollution events since high PM2.5 is rare in training data. To this end, we employed cluster-based undersampling and trained Transformer models for improving extreme event prediction with various cutoff thresholds (12.1 $\mu g/m^3$ and 35.5 $\mu g/m^3$) and partial sampling ratios (10/90, 20/80, 30/70, 40/60, 50/50). Our results show that the 35.5 $\mu g/m^3$ threshold in conjunction with the 20/80 partial sampling ratio yielded the most optimal performance, with an RMSE of 2.128, MAE of 1.418, and R^2 of 0.935, and that it performed well for the prediction of high PM2.5 events. Models trained on resampled data generally performed better than models trained on original data which shows the importance of resampling strategies for increasing the accuracy of air quality prediction for high-pollution levels. These findings provide crucial input to the optimization of air quality forecast models for better prediction of extreme pollution events. Through the enhancement of the ability to predict high concentrations of PM2.5, this study is assisting in the development of more effective public health and environmental policy to mitigate the impacts of air pollution and the mounting dangers of climate change-driven wildfires.

I. Introduction

Air pollution stands as one of the most urgent global health threats, ranking as the second leading risk factor for premature mortality worldwide. In 2021 alone, it was linked to approximately 8.1 million deaths globally, highlighting its devastating health implications [1]. A major contributor to this burden is fine particulate matter (PM2.5) which consists of particles with diameters of 2.5 micrometers or small. Due to their tiny size, PM2.5 particles can infiltrate deep into the respiratory tract and even enter the bloodstream, posing serious health risks. According to the Global Burden of Disease (GBD) study, exposure to ambient PM2.5 resulted in around 4.14 million deaths globally in 2019 [2]. These particles have been associated with various adverse health outcomes, including stroke, ischemic heart disease, chronic obstructive pulmonary disease (COPD), and lung cancer [3-8]. The lungs, in particular, are highly susceptible defenses, increasing vulnerability to respiratory infections [9]. Furthermore, growing research links PM2.5 exposure to neurodegenerative disorders, with evidence suggesting that particles may reach the brain via the olfactory nerve due to their minute size [10].

A concerning trend has emerged in recent years: the rise of PM2.5 emissions from wildfires, which are becoming more frequent and severe due to climate change and evolving land management practices. These wildfire-driven particles can be transported across vast distances, affecting regions far beyond the fire's origin [11]. Since the mid-1980s, wildfires in the western United States have escalated in both frequency and intensity, largely driven by rising temperature and earlier spring snowmelts [12]. Climate projections indicate that wildfire-affected areas in this region may grow by 54% between 2046 and 2055 compared to the 1996-2005 baseline [13]. During major wildfire events, PM2.5 concentrations can surge well beyond safe levels, surpassing U.S. Environmental Protection Agency's (EPA) hazardous threshold of 225.5 $\mu g/m^3$ [14]. These intensifying events emphasize the need for proactive strategies, such as enhanced air quality monitoring, stricter emissions regulations, and robust public health advisories.

Accurate PM2.5 forecasting plays a vital role in safeguarding public health by allowing for timely interventions and reducing exposure to harmful pollution levels. However, forecasting remains difficult due to the complex interplay among atmospheric chemistry, weather variability, and human activities, all of which contribute to rapid and unpredictable fluctuations in pollutant concentrations [15, 16]. Capturing both the temporal dynamics and spatial distribution of PM2.5 is essential for precise exposure and health impact assessments. Forecasting methods range from traditional statistical approaches, like the Autoregressive Integrated Moving Average (ARIMA), to more recent artificial intelligence and machine learning (AI/ML) techniques. Nonlinear models such as Support Vector Regression (SVR) and Artificial Neural Networks (ANNs) have shown potential in modeling these complex datasets [17, 18], though ANNs often fall short in capturing intricate features due to their shallow architectures [19].

Deep Learning (DL) models, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have shown considerable improvements in modeling spatial and temporal patterns [20]. Hybrid CNN-LSTM models have further boosted forecasting

performance, especially for spatiotemporal data [21, 22]. Despite this, DL methods face challenges like vanishing gradients and difficulties modeling long-term dependencies [21]. Transformer architectures, originally developed for Natural Language Processing (NLP) tasks [23], are increasingly being explored for PM2.5 forecasting due to their ability to model long-range dependencies via self attention mechanisms [24]. The Informer model [25] introduced improved temporal embeddings to handle non-stationary and long-sequence time series data but did not account for spatial relationships. To address this, [26] proposed a graph-based Transformer with dynamic spatial attention, and [27] introduced Spacetimeformer, which processes multivariate time series by flattening both spatial and temporal components. Recent innovations like the Sparse Attention-based Transformer (STN) [28] and the SpatioTemporal Transformer [29] continue to improve forecasting accuracy and computational efficiency, particularly for wildfire-prone areas.

A persistent challenge in PM2.5 forecasting is the issue of data imbalance, particularly in predicting high pollution levels. AI/ML models typically perform well in forecasting low to moderate concentrations but often fail to predict high PM2.5 events—those exceeding 60 $\mu g/m^3$—due to their rarity in training datasets [30]. These underrepresented high-concentration events lead to systematic underestimations during extreme pollution episodes [31-25]. Although this imbalance is widely recognized, limited work has been done to directly address it within the context of PM2.5 forecasting [36-38].

One promising solution is data augmentation, which enhances the diversity and richness of the training data by incorporating more representative samples. This improves model generalization and robustness, especially for rare events [39-41]. Oversampling techniques aim to increase the number of minority class samples, such as through simple duplication of synthetic methods like SMOTE and ADASYN [42-45]. However, random oversampling can result in overfitting, while synthetic methods strive to balance the dataset more effectively. Undersampling, on the other hand, reduces the size of the majority class. Although random undersampling may lead to loss of information [42], combining it with clustering methods such as k-means clustering can help retain data structure while still achieving balance [46, 47].

Several studies have explored these augmentation methods in air pollution modeling. For instance [48] used random oversampling and SMOTE in combination with an AugResNet model to improve the prediction of high PM2.5 levels. While this approach yielded better results for high concentrations, it focused on retrieval rather than forecasting and used a single cutoff, limiting its broader application. In another study, [49] applied LSTM, GRU, and hybrid models with linear interpolation to enlarge datasets, though this method did not directly address class imbalance, potentially causing overfitting. Meanwhile, [50] addressed dataset bias between urban and rural environments using multiple imputation but did not focus on forecasting extreme PM2.5 levels, making it application context-specific.

Despite progress critical gaps remain. First, the persistent imbalance in datasets limits forecasting accuracy during extreme pollution events, such as those caused by wildfires [31-33]. Second, although augmentation methods like SMOTE and ADASYN have proven useful, their

application specifically for PM2.5 forecasting remains limited and often does not focus on high-concentration scenarios [42, 48, 49]. Lastly, Transformer models—despite their strong potential for long-range modeling and spatiotemporal leaning—have not been fully explored for urban PM2.5 forecasting during extreme events [25, 29].

This study seeks to bridge these gaps by applying cluster-based undersampling with various majority-to-minority ratios to better represent extreme PM2.5 events in training datasets. The thresholds used to define majority and minority classes are based on two EPA standards for air quality, ensuring real-world relevance. The study employs a Transformer-based model with multi-head sparse attention, drawing on advancements from models like Informer and Spacetimeformer to effectively capture long-term dependencies and spatiotemporal variability.

II. Background

This study focuses on three major urban centers in the northeastern United States: New York City, Philadelphia, and Washington, D.C. These cities are densely populated and subject to heavy traffic and industrial activity, all of which contribute to elevated PM2.5 levels. Urban environments are particularly susceptible to fine particulate pollution due to emission from vehicles, manufacturing, and residential heating systems, making the key targets for air quality surveillance and predictive modeling [51, 52]. Forecasting air quality in these regions is especially complex due to the intricate interactions between local emissions sources and broader atmospheric transport processes.

A major air quality event occurred in 2023 when extensive Canadian wildfires severely affected pollution levels across North America, including the northeastern U.S. [53, 54]. Massive plumes of smoke and fine particulate matter from the fires were carried southward by prevailing winds, causing dramatic surges in PM2.5 levels in cities like New York, Philadelphia, and Washington D.C. [55]. During this period, PM2.5 concentrations reached hazardous levels, significantly impairing visibility and prompting urgent public health advisories [56]. Accurate PM2.5 forecasts are essential in these urban settings, particularly during extreme events such as the 2023 wildfires, to inform public health responses and reduce exposure risks [57].

Variable	Source	Unit
PM2.5	EPA AirNow	$\mu g/m^3$
AOD	MODIS MAIAC (Terra and Aqua)	
Boundary Layer Height (BLH)	ECMWF ERA5-hourly	meter
Relative Humidity	ECMWF ERA5-hourly	%

Temperature (at 2m)	ECMWF ERA5-hourly	K
Surface Pressure	ECMWF ERA5-hourly	Pa
Wind Speed	ECMWF ERA5-hourly	m/s
Elevation	USGS	meter

Table 1. Covariates and Data Sources

The variables used for PM2.5 forecasting are summarized in Table 1. The target variable, surface-level PM2.5 was sourced from the EPA's AirNow system, which predictor variables included aerosol optical depth (AOD), meteorological data, and elevation. AOD data was obtained from the MODIS instrument using the MAIAC algorithm, which provides a proxy for atmospheric particulate content closely correlated with surface PM2.5 [58, 59]. Meteorological variables known to influence PM2.5 dynamics—such as mixing height, humidity, temperature, surface pressure, and wind speed—were retrieved from the ERA5 dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) {58, 60]. These parameters influence particle formation, dispersion, and chemical transformation. Elevation, another key geographic covariate, was included using data from the U.S. Geological Survey (USGS), given its role in modulating wind flow and pollutant accumulation [63].

Hourly PM2.5 concentrations were collected from the AirNow platform, maintained by the U.S EPA. AirNow compiles near-real-time air quality readings from a network of over 120 federal, state, tribal, provincial, and local agencies. Data from monitoring stations in New York, Philadelphia, and Washington, D.C. were obtained via the AirNow API. All measurements undergo rigorous quality assurance protocols before public release. AirNow was chosen over the Air Quality System (AQS) due to its more timely availability, which is critical for forecasting applications. These ground-level measurements served as the baseline for model evaluation and validation in the selected urban regions.

AOD values were derived from the MODIS instruments aboard NASA's Terra and Aqua satellites, using the MAIAC (Multi-Angle Implementation of Atmospheric Correction) algorithm. These satellites provide daily AOD products with a 1 km x 1 km spatial resolution, collected at around 10:30 a.m. (Terra) and 1:30 p.m. (Aqua) local time [64, 65]. MAIAC is a sophisticated retrieval method designed to perform well over both dark vegetated and bright desert surfaces, making it highly applicable for air quality studies due to its fine resolution and broad coverage [66, 67]. The MCD19A2 Version 6 AOD product at 550 nm was used, and only values flagged as "best quality" were retained to ensure data reliability. The dataset was obtained from the Level-1 and Atmosphere Archive and Distribution System Distributed Active Archive Center (LAADS DAAC) [68].

Meteorological predictors were drawn from the ERA5 reanalysis dataset developed by ECMWF and distributed via the Copernicus Climate Data Store (C3S) ERA5 offers hourly global data since 1940, with a spatial resolution of 0.25° × 0.25° [69, 70]. This makes it suitable for both real-time and retrospective analysis. For this study, key variables included boundary layer height (BLH), relative humidity, surface pressure (SP), 2-meter temperature (T2M), and 10-meter wind speed. Wind speed was calculated from the eastward (U) and northward (V) wind components using the Euclidian norm. These factors directly influence PM2.5 concentration through mechanisms like vertical dispersion, chemical formation, and particle growth [60-62].

Elevation data were sourced from the Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010) dataset, which provides topographic information at multiple spatial resolutions [71]. This global dataset combines data from various sources and applies subsampling techniques to ensure broad coverage and consistency. For this study, elevation data at 30 arc-seconds (~1 km) resolution were downloaded from the USGS repository. Elevation plays a crucial role in pollution modeling as it affects atmospheric mixing and pollutant transport; higher altitudes often experience better dispersion due to stronger wind speeds and lower pressure conditions.

III. Applications

This research has significant real-world applications in public health, environmental monitoring, and emergency response planning. By enhancing the ability to accurately forecast PM2.5 concentrations—particularly during extreme pollution events such as wildfires—this study supports more timely and effective health advisories. Early warnings based on improved forecasts allow vulnerable populations such as children, the elderly, and those with respiratory conditions, to take precautionary measures to minimize exposure. Public health agencies can use these predictions to issue targeted alerts, distribute protective equipment like masks and air purifiers, and manage hospital readiness during air quality crises. Moreover, city planners and policymakers can integrate this forecasting framework into long-term strategies for urban sustainability, leveraging data-driven insights to regulate emissions and design healthier, more resilient communities.

Beyond public health, the findings also offer substantial value for advancing environmental justice and climate resilience. Urban neighborhoods with limited access to healthcare and clean air resources are often disproportionately affected by air pollution. Accurate, localized PM2.5 forecasts can help prioritize interventions in these high-risk communities, ensuring more equitable distribution of environmental protections. Furthermore, as climate change continues to drive the frequency and intensity of wildfires, this research contributes to a broader toolkit for climate adaptation. The integration of Transformer-based models and data augmentation techniques enhances the robustness of air quality predictions under non-stationary and rapidly evolving conditions, making it a powerful resource for governments, environmental

organizations, and disaster preparedness agencies working to mitigate the impacts of worsening air pollution trends.

IV. Methods:

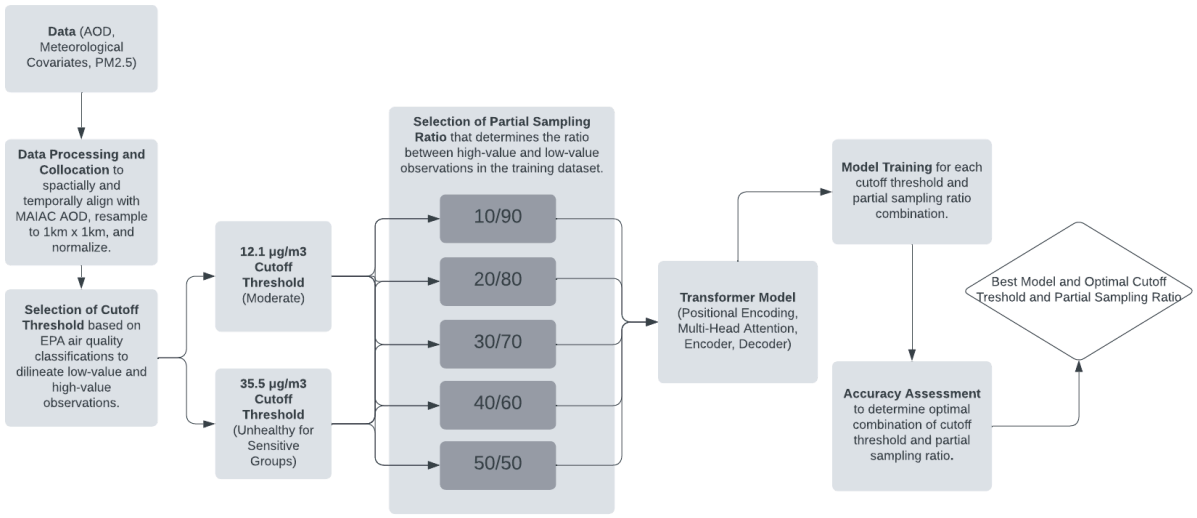

Figure 1. Workflow Diagram

This study aims to forecast PM2.5 concentrations in urban areas by applying Transformer-based deep learning and cluster-based undersampling to mitigate data imbalance. Our input includes spatiotemporally aligned data from the AirNow PM2.5 sensor network, MODIS satellite-derived aerosol optical depth (AOD), meteorological data from ERA5, and elevation data from GMTED2010. The out is a time series of predicted PM2.5 values across New York City, Philadelphia, and Washington, D.C. To produce accurate forecasts, we preprocess and align all datasets temporally and spatially, normalize the input features, apply class-based data augmentation via undersampling, and train a Transformer model on the resulting datasets.

Each dataset was matched to the MODIS Terra and Aqua over pass times (~10:30 and 13:30 local time), and hourly measurements were averaged accordingly. Meteorological variables (originally at 0.25° resolution) were interpolated to 1 km x 1 km grids and reprojected to the USA Contiguous Lambert Conformal Conic system using ArcGIS Pro 3.2.1. PM2.5 point data were converted into gridded surfaces via kriging. Min-max normalization was then applied to rescale inputs to the [0, 1] range.

A previous study [48] adopted 75 µg/m3 as a cutoff based on China's PM2.5 air quality guidelines. In contrast, we chose thresholds aligned with U.S. EPA standards: 12.1 µg/m3 (Moderate) and 35.5 µg/m3 (Unhealthy for Sensitive Groups) to define high- and low-value PM2.5 categories. This dual-threshold approach, unlike earlier studies used a single fixed threshold, enables sensitivity analysis of model performance across different pollution levels.

For the dataset of 4,026,240 valid entries, the 12.1 μg/m3 threshold classified ~3.47 million as low and ~0.55 million as high. The 35.5 μg/m3 threshold classified ~4 million as low and ~22,000 as high.

To handle severe class imbalance, we implemented k-means-based undersampling. The majority class was clustered to preserve structural integrity, and representative points were selectively sampled from each cluster to balance the dataset without sacrificing diversity. This contrasts with traditional oversampling techniques (e.g. SMOTE or random duplication) which can induce overfitting. We reserved 20% of the dataset for consistent testing, while 80% was used for training with different sampling ratios, ensuring consistent evaluation across models. This approach allowed us to examine multiple partial samplings (10/90, 20/80, 30/70, 40/60, 50/50), with final training datasets containing approximately 35,000 instances each.

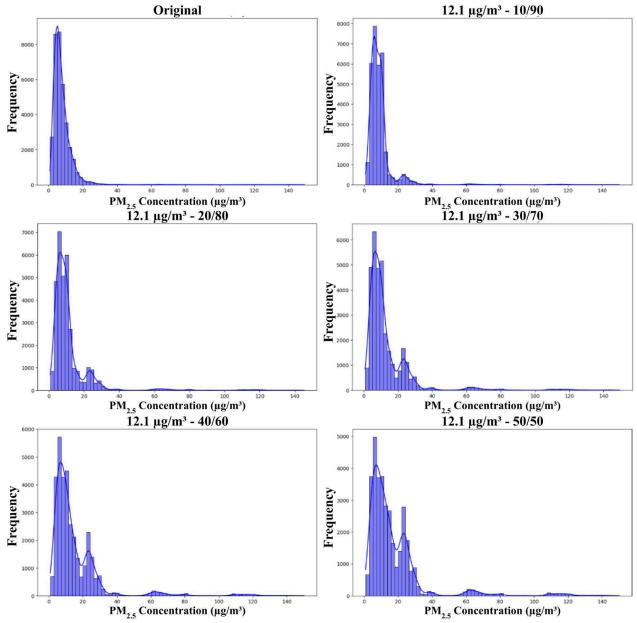

Figure 2. Distribution of training dataset with cutoff threshold 12.1 μg/m3 and partial sampling ratio, from top left to bottom right, of none (original distribution), 10/90, 20/80, 30/70, 40/60, 50/50.

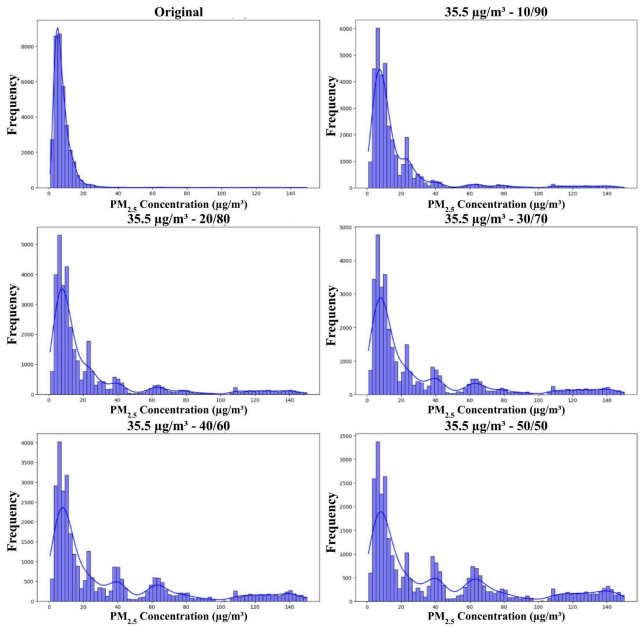

Figure 3. Distribution of training dataset with cutoff threshold 35.5 µg/m3 and partial sampling ratio, from top left to bottom right, of none (original distribution), 10/90, 20/80, 30/70, 40/60, 50/50.

Rather than enforcing a strict 1:1 class balance, we explored partial sampling ratios informed by prior literature [72, 73], which suggests that a 0.75 minority-to-majority ratio often yields optimal performance. This allowed us to test more realistic class balances that reflect natural pollution variability while improving model generalization. Although training sets had the same size, internal distributions varied depending on the chosen cutoff and ratio.

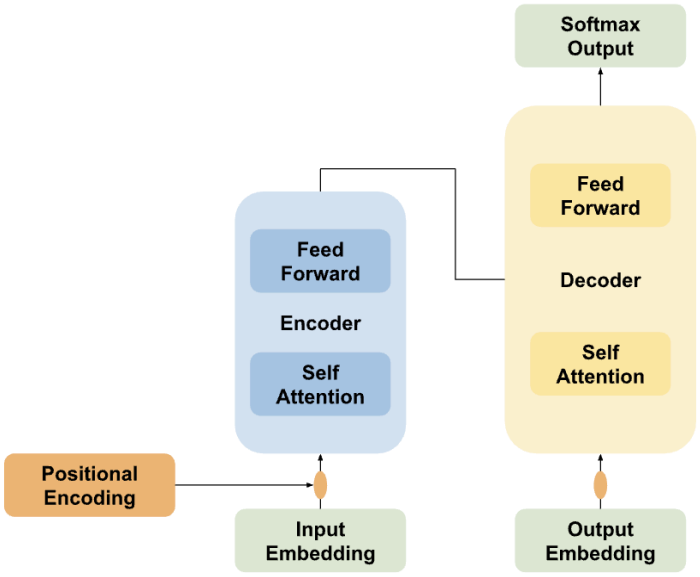

Figure 4. Transformer Architecture

Our model incorporates a Transformer with six encoder and six decoder layers, each utilizing multi-head attention, residual connections, layer normalization, and feed-forward networks. A positional encoding mechanism embed temporal information using sinusoidal function. Self-attention quantifies the dependencies between different parts of the input time series and calculates how much of the information corresponding to a time slice should be focused on in future forecasts. This architecture is particularly well-suited for modeling long-term dependencies and spatiotemporal patterns in PM2.5 dynamics. Our design builds on advancements from Informer and Spacetimeformer. A simplified visualization of the architecture is shown in Figure 4.

We trained the models using the Adam optimizer. Hyperparameters such as learning rate and dropout were tuned manually through iterative experimentation. We maintained consistent hyperparameters settings across experiments to ensure fair model comparison.

V. **Results**:

12.1 ug/m3 Cutoff Threshold						
Partial Sampling Ratio	Whole Dataset			High Values		
	RMSE	MAE	R-squared	RMSE	MAE	R-squared
Original	3.247 ± 0.101	0.676 ± 0.021	0.819 ± 0.025	32.748 ± 1.014	27.318 ± 0.846	0.037 ± 0.001
10/90	3.291 ± 0.102	0.743 ± 0.023	0.743 ± 0.023	30.040 ± 0.930	20.750 ± 0.642	0.192 ± 0.006
20/80	3.161 ± 0.098	1.171 ± 0.036	0.831 ± 0.026	26.544 ± 0.822	19.481 ± 0.603	0.374 ± 0.012
30/70	2.888 ± 0.089	1.570 ± 0.049	0.862 ± 0.027	25.822 ± 0.800	19.259 ± 0.596	0.409 ± 0.013
40/60	2.881 ± 0.089	1.355 ± 0.042	0.862 ± 0.027	23.818 ± 0.737	17.782 ± 0.551	0.501 ± 0.016
50/50	2.820 ± 0.087	1.068 ± 0.033	0.870 ± 0.027	21.776 ± 0.674	14.438 ± 0.447	0.587 ± 0.018

Table 2. Accuracy measurements of models trained on data augmented with cutoff threshold 12.1 µg/m3 and different partial sampling ratios tested on the whole and high-value testing dataset.

The most effective partials sampling ratios were identified by evaluation model performance metrics across both augmented and original datasets for each cutoff threshold. For models trained using the 12.1 µg/m³ cutoff, the results in Table 2 indicate that as the sampling becomes more balanced (from 10/90 to 50/50), both RMSE and MAE values generally decline, reflecting improved model accuracy. The 50/50 ratio yielded the best overall results, with an RMSE of 2.820, MAE of 1.068, and an R^2 value of 0.870. This high R^2 suggests a strong alignment between the predicted and observed PM2.5 values, making this configuration the most effective for datasets where a balanced class distribution is desired.

35.5 ug/m3 Cutoff Threshold						
Partial Sampling Ratio	Whole Dataset			High Values		
	RMSE	MAE	R-squared	RMSE	MAE	R-squared
Original	3.247 ± 0.103	1.629 ± 0.050	0.819 ± 0.025	32.748 ± 1.014	27.318 ± 0.846	0.037 ± 0.001
10/90	2.334 ± 0.072	1.629 ± 0.050	0.918 ± 0.028	20.200 ± 0.625	14.127 ± 0.437	0.648 ± 0.020
20/80	2.128 ± 0.066	1.418 ± 0.044	0.935 ± 0.029	15.705 ± 0.486	10.308 ± 0.319	0.796 ± 0.025
30/70	2.359 ± 0.073	1.709 ± 0.053	0.916 ± 0.028	16.464 ± 0.510	12.484 ± 0.387	0.773 ± 0.024
40/60	2.479 ± 0.077	1.766 ± 0.055	0.904 ± 0.028	16.936 ± 0.524	13.213 ± 0.409	0.758 ± 0.023
50/50	2.738 ± 0.085	1.918 ± 0.059	0.878 ± 0.027	19.555 ± 0.605	14.650 ± 0.454	0.671 ± 0.021

Table 3. Accuracy measurements of models trained on data augmented with cutoff threshold 35.5 µg/m3 and different partial sampling ratios tested on the whole and high-value testing dataset.

For experiments using the 35.5 µg/m³ cutoff, results are presented in Table 3. Here, the optimal sampling ratio shifted to 20/80, which produced the lowest RMSE (2.128), MAE (1.418) and the highest R^2 (0.935). This combination effectively balances the representation of both majority and

minority class instances while minimising prediction error. Notably, the same 20/80 partial sampling ratio also performed best on high PM2.5 values, achieving an RMSE of 15.705, MAE of 10.308, and an R^2 of 0.796. These outcomes highlight the 20/80 ratio's suitability for forecasting extreme pollution events.

When comparing models trained on the unaltered dataset to those trained with the resampled data, the original dataset consistently showed inferior performance, particularly in RMSE and R^2. This reinforced the importance of data augmentation in enhancing forecast accuracy.

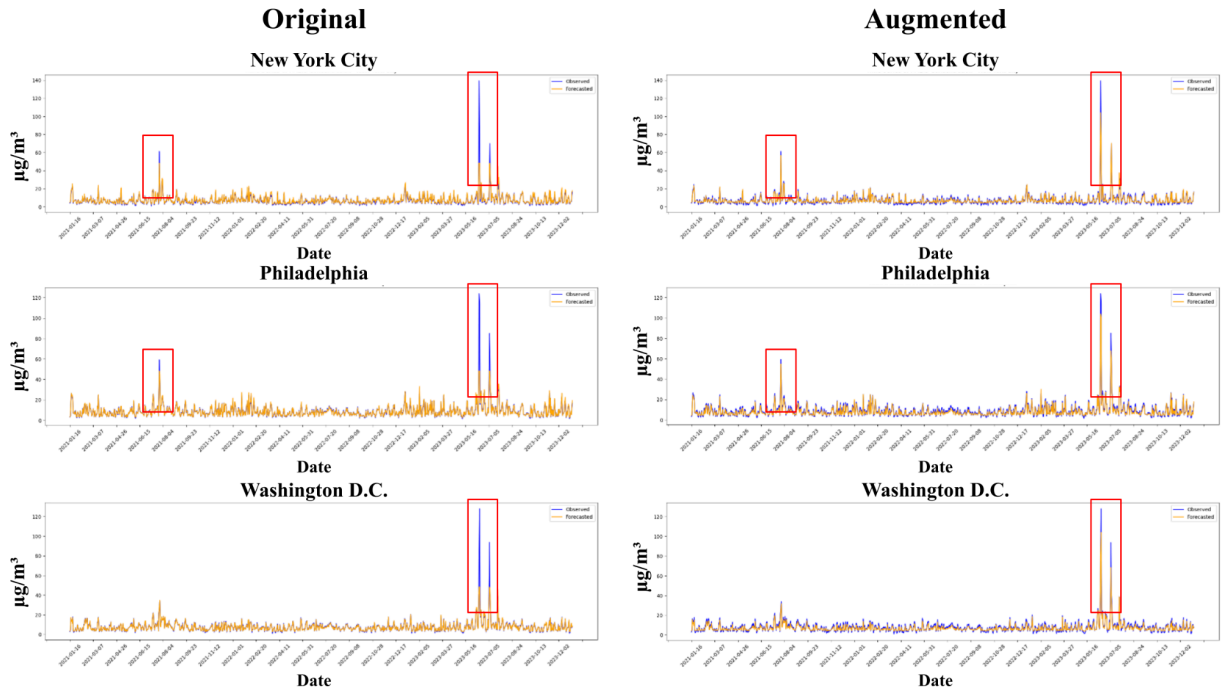

Figure 5. Time series of observed (blue) and forecast (orange) PM2.5 concentrations from 2021 to 2023 in New York City (top), Philadelphia (middle), and Washington D.C. (bottom). The left column shows predictions from the model trained on original distribution and the right column shows predictions from the model trained on data treated with optimal data augmentation, determined in Section 4.1. The red-boxed regions highlight extreme PM2.5 events.

Figure 5 visualized the time series comparison between observed and predicted PM2.5 levels, focusing on models trained with and without augmentation at the 35.5 µg/m³ cutoff and a 20/80 sampling ratio. In all three cities, the model trained on the original dataset performs well at forecasting low PM2.5 concentrations, demonstrating a close match between predicted and actual values in these lower ranges. However, it fails to accurately capture higher pollution levels, flattening out in response to extreme events, as seen in the red-highlighted sections. This shortfall results from the skewed data distribution, where low PM2.5 levels dominate, limiting the model's ability to generalize to rare high-value events.

Conversely, the model trained on the 35.5 µg/m³ and 20/80 partial sampling ratio augmented dataset performs much better at identifying and predicting high PM2.5 concentrations. While this

model shows slightly reduced accuracy for lower PM2.5 levels, it yields significantly stronger RMSE and R^2 values overall. In the red-boxed regions, the predicted peaks closely track the observed data, illustrating the model's enhanced ability to forecast severe pollution episodes. Although MAE increases slightly due to the introduction of greater variability and minor errors, RMSE—more sensitive to large deviations—improves, indicating stronger performance where it matters most. This suggests that the augmented model is more robust for forecasting critical high-concentration pollution events.

In summary, the key distinction between the two models lies in their focus: the model trained on unbalanced data excels at predicting routine, low-level PM2.5, while the augmented model prioritizes high-concentration events at a minor cost to low-level accuracy. Given the importance of anticipating extreme pollution episodes, the trade-off in accuracy is worthwhile, especially as demonstrated by RMSE improvements. Overall, the augmented model is clearly more effective at capturing the full range of PM2.5 behaviors, particularly those with the greatest public health implications.

VI. Limitations

While this study demonstrates the effectiveness of Transformer models and data augmentation techniques for PM2.5 forecasting, several limitations remain. First, the model was trained and evaluated using data from only three northeastern U.S. cities, which restricts its generalizability. Application to other regions would require retraining to account for differences in local pollution sources, weather patterns, and population density. Second, although cluster-based undersampling improved model performance on high-value PM2.5 points, it does not perfectly preserve the characteristics of the original dataset potentially introducing new biases. Third, there is a trade-off between overall accuracy and the model's ability to predict extreme pollution levels; improvements in high-value forecasts can sometimes come at the costs of lower accuracy for common, lower pollution values. Finally, expanding to multistep prediction introduces challenges with error propagation, as forecast inaccuracies at each step can accumulate and degrade performance over longer time horizons.

VII. Conclusion

This study shows that the 35.5 μg/m³ threshold consistently delivered stronger model performance than the 12.1 μg/m³ threshold, particularly for RMSE and R^2 metrics, due to its better representation of high-pollution events. The choice of partial sampling ratio was equally important. While 50/50 worked best at the lower threshold, a 20/80 split proved optimal at the higher one. The top-performing model (RMSE: 2.128, MAE: 1.418, R^2: 0.935) used the 35.5 μg/m³ cutoff with 20/80 sampling. Overall, models trained with resampled data outperformed those using unaugmented datasets, reinforcing the effectiveness of data augmentation in addressing class imbalance and improving prediction accuracy, especially for extreme pollution scenarios.

These findings hold valuable practical implications. Accurate PM2.5 forecasting is essential for guiding public health decisions in urban areas exposed to hazardous air quality. Customizing threshold selection and resampling strategies to the characteristics of the dataset leads to more reliable predictions, which can inform policy and emergency interventions.

VIII. Future Work

Future research could benefit from incorporating additional data sources relevant to PM2.5 fluctuations, such as traffic density, industrial emissions, and weather forecasts. Expanding the model to forecast multiple future time steps (multistep forecasting) would also increase its utility for public health planning. Moreover, techniques like Generative Adversarial Networks (GANs) could be explored to synthetically generate extreme pollution scenarios for training, a crucial step for improving rare event predictions. Extending this framework to other pollutants like NO2, SO2, and O3 would facilitate comprehensive air quality modeling. Finally, applying this approach across diverse geographic regions would test the model's generalizability and may require customizing the sampling strategy to local pollution dynamics and meteorology.

References

1. State of Global Air Report. Available online: **https://www.stateofglobalair.org/resources/report/state-global-air-report-2024** (accessed on 16 January 2025).

2. McDuffie, E.; Martin, R.; Yin, H.; Brauer, M. Global Burden of Disease from Major Air Pollution Sources (GBD MAPS): A Global Approach. *Res. Rep. Health Eff. Inst.* **2021**, *2021*, 1–45.

3. Gao, X.; Koutrakis, P.; Coull, B.; Lin, X.; Vokonas, P.; Schwartz, J.; Baccarelli, A.A. Short-term exposure to PM2.5 components and renal health: Findings from the Veterans Affairs Normative Aging Study. *J. Hazard. Mater.* **2021**, *420*, 126557.

4. Gilcrease, G.W.; Padovan, D.; Heffler, E.; Peano, C.; Massaglia, S.; Roccatello, D.; Radin, M.; Cuadrado, M.J.; Sciascia, S. Is air pollution affecting the disease activity in patients with systemic lupus erythematosus? State of the art and a systematic literature review. *Eur. J. Rheumatol.* **2020**, *7*, 31.

5. Hystad, P.; Larkin, A.; Rangarajan, S.; AlHabib, K.F.; Avezum, Á.; Calik, K.B.T.; Chifamba, J.; Dans, A.; Diaz, R.; du Plessis, J.L.; et al. Associations of outdoor fine particulate air pollution and cardiovascular disease in 157 436 individuals from 21 high-income, middle-income, and low-income countries (PURE): A prospective cohort study. *Lancet Planet. Health* **2020**, *4*, e235–e245.

6. Lao, X.Q.; Guo, C.; La-yun, C.; Bo, Y.; Zhang, Z.; Chuang, Y.C.; Jiang, W.K.; Lin, C.; Tam, T.; Lau, A.K.H.; et al. Long-term exposure to ambient fine particulate matter (PM 2.5) and incident type 2 diabetes: A longitudinal cohort study. *Diabetologia* **2019**, *62*, 759–769.

7. Liu, L.; Zhang, Y.; Yang, Z.; Luo, S.; Zhang, Y. Long-term exposure to fine particulate constituents and cardiovascular diseases in Chinese adults. *J. Hazard. Mater.* **2021**, *416*, 126051.

8. Thangavel, P.; Park, D.; Lee, Y.C. Recent Insights into Particulate Matter (PM2.5)-Mediated Toxicity in Humans: An Overview. *Int. J. Environ. Res. Public Health* **2022**, *19*, 7511.

9. Jia, H.; Liu, Y.; Guo, D.; He, W.; Zhao, L.; Xia, S. PM2.5-induced pulmonary inflammation via activating of the NLRP3/caspase-1 signaling pathway. *Environ. Toxicol.* **2021**, *36*, 298–307.

10. Lee, S.; Lee, W.; Kim, D.; Kim, E.; Myung, W.; Kim, S.Y.; Kim, H. Short-term PM2.5 exposure and emergency hospital admissions for mental disease. *Environ. Res.* **2019**, *171*, 313–320.

11. Sharma, A.; Valdes, A.C.F.; Lee, Y. Impact of Wildfires on Meteorology and Air Quality (PM2.5 and O3) over Western United States during September 2017. *Atmosphere* **2022**, *13*, 262.

12. Westerling, A.L.; Hidalgo, H.G.; Cayan, D.R.; Swetnam, T.W. Warming and earlier spring increase Western U.S. forest wildfire activity. *Science* **2006**, *313*, 940–943.

13. Spracklen, D.V.; Mickley, L.J.; Logan, J.A.; Hudman, R.C.; Yevich, R.; Flannigan, M.D.; Westerling, A.L. Impacts of climate change from 2000 to 2050 on wildfire activity and carbonaceous aerosol concentrations in the western United States. *J. Geophys. Res. Atmos.* **2009**, *114*, 20301.

14. EPA AQI. Final Updates to the Air Quality Index (AQI) for Particulate Matter—Fact Sheet and Common Questions. Available online: **https://www.epa.gov/system/files/documents/2024-02/pm-naaqs-air-quality-index-fact-sheet.pdf** (accessed on 16 January 2025).

15. Liou, N.-C.; Luo, C.-H.; Mahajan, S.; Chen, L.-J. Why is Short-Time PM2.5 Forecast Difficult? The Effects of Sudden Events. *IEEE Access* **2020**, *8*, 12662–12674.

16. Ma, Z.; Dey, S.; Christopher, S.; Liu, R.; Bi, J.; Balyan, P.; Liu, Y. A review of statistical methods used for developing large-scale and long-term PM2.5 models from satellite data. *Remote Sens. Environ.* **2022**, *269*, 112827.

17. Abedi, A.; Baygi, M.M.; Poursafa, P.; Mehrara, M.; Amin, M.M.; Hemami, F.; Zarean, M. Air pollution and hospitalization: An autoregressive distributed lag (ARDL) approach. *Environ. Sci. Pollut. Res.* **2020**, *27*, 30673–30680.

18. Agarwal, S.; Sharma, S.; Suresh, R.; Rahman, M.H.; Vranckx, S.; Maiheu, B.; Blyth, L.; Janssen, S.; Gargava, P.; Shukla, V.K.; et al. Air quality forecasting using artificial neural networks with real time dynamic error correction in highly polluted regions. *Sci. Total Environ.* **2020**, *735*, 139454.

19. Ding, W.; Zhang, J.; Leung, Y. Prediction of air pollutant concentration based on sparse response back-propagation training feedforward neural networks. *Environ. Sci. Pollut. Res.* **2016**, *23*, 19481–19494.

20. Gao, X.; Li, W. A graph-based LSTM model for PM2.5 forecasting. *Atmos. Pollut. Res.* **2021**, *12*, 101150.

21. Wen, C.; Liu, S.; Yao, X.; Peng, L.; Li, X.; Hu, Y.; Chi, T. A novel spatiotemporal convolutional long short-term neural network for air pollution prediction. *Sci. Total Environ.* **2019**, *654*, 1091–1099.

22. Zhang, Z.; Zeng, Y.; Yan, K. A hybrid deep learning technology for PM2.5 air quality forecasting. *Environ. Sci. Pollut. Res.* **2021**, *28*, 39409–39422.

23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.

24. Dong, J.; Zhang, Y.; Hu, J. Short-term air quality prediction based on EMD-transformer-BiLSTM. *Sci. Rep.* **2024**, *14*, 20513.

25. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 11106–11115.

26. Li, Y.; Moura, J.M.F. Forecaster: A Graph Transformer for Forecasting Spatial and Time-Dependent Data. *Front. Artif. Intell. Appl.* **2020**, *325*, 1293–1300.

27. Grigsby, J.; Wang, Z.; Nguyen, N.; Qi, Y. Long-Range Transformers for Dynamic Spatiotemporal Forecasting. *arXiv* **2021**, arXiv:2109.12218v3.

28. Zhang, Z.; Zhang, S. Modeling air quality PM2.5 forecasting using deep sparse attention-based transformer networks. *Int. J. Environ. Sci. Technol.* **2023**, *20*, 13535–13550.

29. Yu, M.; Masrur, A.; Blaszczak-Boxe, C. Predicting hourly PM2.5 concentrations in wildfire-prone areas using a SpatioTemporal Transformer model. *Sci. Total Environ.* **2023**, *860*, 160446.

30. Yan, X.; Zang, Z.; Jiang, Y.; Shi, W.; Guo, Y.; Li, D.; Zhao, C.; Husi, L. A Spatial-Temporal Interpretable Deep Learning Model for improving interpretability and predictive accuracy of satellite-based PM2.5. *Environ. Pollut.* **2021**, *273*, 116459.

31. Li, T.; Shen, H.; Yuan, Q.; Zhang, X.; Zhang, L. Estimating Ground-Level PM2.5 by Fusing Satellite and Station Observations: A Geo-Intelligent Deep Learning Approach. *Geophys. Res. Lett.* **2017**, *44*, 11985–11993.

32. Liu, J.; Weng, F.; Li, Z. Ultrahigh-Resolution (250 m) Regional Surface PM2.5 Concentrations Derived First from MODIS Measurements. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12.

33. Ma, Z.; Hu, X.; Huang, L.; Bi, J.; Liu, Y. Estimating ground-level PM2.5 in china using satellite remote sensing. *Environ. Sci. Technol.* **2014**, *48*, 7436–7444.

34. Xu, Y.; Ho, H.C.; Wong, M.S.; Deng, C.; Shi, Y.; Chan, T.C.; Knudby, A. Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level PM2.5. *Environ. Pollut.* **2018**, *242*, 1417–1426. [**PubMed**]

35. Zhan, Y.; Luo, Y.; Deng, X.; Chen, H.; Grieneisen, M.L.; Shen, X.; Zhu, L.; Zhang, M. Spatiotemporal prediction of continuous daily PM2.5 concentrations across China using a spatially explicit machine learning algorithm. *Atmos. Environ.* **2017**, *155*, 129–139.

36. Lu, Y.; Giuliano, G.; Habre, R. Estimating hourly PM2.5 concentrations at the neighborhood scale using a low-cost air sensor network: A Los Angeles case study. *Environ. Res.* **2021**, *195*, 110653.

37. Xiao, Q.; Zheng, Y.; Geng, G.; Chen, C.; Huang, X.; Che, H.; Zhang, X.; He, K.; Zhang, Q. Separating emission and meteorological contributions to long-term PM2.5trends over eastern China during 2000–2018. *Atmos. Chem. Phys.* **2021**, *21*, 9475–9496.

38. Zhang, S.; Mi, T.; Wu, Q.; Luo, Y.; Grieneisen, M.L.; Shi, G.; Yang, F.; Zhan, Y. A data-augmentation approach to deriving long-term surface SO2 across Northern China: Implications for interpretable machine learning. *Sci. Total Environ.* **2022**, *827*, 154278.

39. Feng, W.; Boukir, S.; Huang, W. Margin-Based Random Forest for Imbalanced Land Cover Classification. In Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS), Yokohama, Japan, 28 July–2 August 2019; pp. 3085–3088.

40. Stivaktakis, R.; Tsagkatakis, G.; Tsakalides, P. Deep Learning for Multilabel Land Cover Scene Categorization Using Data Augmentation. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1031–1035.

41. Yu, X.; Wu, X.; Luo, C.; Ren, P. Deep learning in remote sensing scene classification: A data augmentation enhanced convolutional neural network framework. *GI Sci. Remote Sens.* **2017**, *54*, 741–758.

42. Mohammed, R.; Rawashdeh, J.; Abdullah, M. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. In Proceedings of the 2020 11th International Conference on Information and Communication Systems ICICS, Irbid, Jordan, 7–9 April 2020; pp. 243–248.

43. Khan, A.A.; Chaudhari, O.; Chandra, R. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Syst. Appl.* **2024**, *244*, 122778.

44. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357.

45. Torgo, L.; Ribeiro, R.P.; Pfahringer, B.; Branco, P. SMOTE for Regression. In *Progress in Artificial Intelligence. EPIA 2013. Lecture Notes in Computer Science*; Correia, L., Reis, L.P., Cascalho, J., Eds.; Springer: Berlin, Heidelberg, 2013; Volume 8154.

46. Lin, W.C.; Tsai, C.F.; Hu, Y.H.; Jhang, J.S. Clustering-based undersampling in class-imbalanced data. *Inf. Sci.* **2017**, *409–410*, 17–26.

47. Yen, S.J.; Lee, Y.S. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Syst. Appl.* **2009**, *36*, 5718–5727.

48. Yin, S.; Li, T.; Cheng, X.; Wu, J. Remote sensing estimation of surface PM2.5 concentrations using a deep learning model improved by data augmentation and a particle size constraint. *Atmos. Environ.* **2022**, *287*, 119282.

49. Flores, A.; Valeriano-Zapana, J.; Yana-Mamani, V.; Tito-Chura, H. PM2.5 prediction with Recurrent Neural Networks and Data Augmentation. In Proceedings of the 2021 IEEE Latin American Conference on Computational Intelligence LA-CCI, Temuco, Chile, 2–4 November 2021.

50. Mi, T.; Tang, D.; Fu, J.; Zeng, W.; Grieneisen, M.L.; Zhou, Z.; Jia, F.; Yang, F.; Zhan, Y. Data augmentation for bias correction in mapping PM2.5 based on satellite retrievals and ground observations. *Geosci. Front.* **2024**, *15*, 101686.

51. Kloog, I.; Chudnovsky, A.A.; Just, A.C.; Nordio, F.; Koutrakis, P.; Coull, B.A.; Lyapustin, A.; Wang, Y.; Schwartz, J. A new hybrid spatio-temporal model for estimating daily multi-year PM2.5 concentrations across northeastern USA using high resolution aerosol optical depth data. *Atmos. Environ.* **2014**, *95*, 581–590.

52. Qin, Y.; Kim, E.; Hopke, P.K. The concentrations and sources of PM2.5 in metropolitan New York City. *Atmos. Environ.* **2006**, *40* (Suppl. S2), 312–332.

53. Wang, Z.; Wang, Z.; Zou, Z.; Chen, X.; Wu, H.; Wang, W.; Su, H.; Li, F.; Xu, W.; Liu, Z.; et al. Severe Global Environmental Issues Caused by Canada's Record-Breaking Wildfires in 2023. *Adv. Atmos. Sci.* **2024**, *41*, 565–571.

54. Yu, M.; Zhang, S.; Ning, H.; Li, Z.; Zhang, K. Assessing the 2023 Canadian wildfire smoke impact in Northeastern US: Air quality, exposure and environmental justice. *Sci. Total Environ.* **2024**, *926*, 171853.

55. Bella, T. Philadelphia's hazardous air quality from Canadian wildfires is worst level in city since 1999—The Washington Post. *The Washington Post*, 8 June 2023. Available online: **https://www.washingtonpost.com/climate-environment/2023/06/08/philadelphia-air-quality-worst-wildfire-smoke/** (accessed on 16 January 2025).

56. Deegan, D. Canadian Wildfires Prompt Poor Air Quality Alert for Parts of New England on 7 June 2023. *US EPA*, 7 June 2023. Available online: **https://www.epa.gov/newsreleases/canadian-wildfires-prompt-poor-air-quality-alert-parts-new-england-june-7-2023** (accessed on 16 January 2025).

57. Xu, Y.; Yang, W.; Wang, J. Air quality early-warning system for cities in China. *Atmos. Environ.* **2017**, *148*, 239–257.

58. Huang, F.; Li, X.; Wang, C.; Xu, Q.; Wang, W.; Luo, Y.; Tao, L.; Gao, Q.; Guo, J.; Chen, S.; et al. PM2.5 Spatiotemporal Variations and the Relationship with Meteorological Factors during 2013–2014 in Beijing, China. *PLoS ONE* **2015**, *10*, e0141642.

59. Yang, Z.; Zdanski, C.; Farkas, D.; Bang, J.; Williams, H. Evaluation of Aerosol Optical Depth (AOD) and PM2.5 associations for air quality assessment. *Remote Sens. Appl. Soc. Environ.* **2020**, *20*, 100396.

60. Chen, Z.; Chen, D.; Zhao, C.; Kwan, M.; Cai, J.; Zhuang, Y.; Zhao, B.; Wang, X.; Chen, B.; Yang, J.; et al. Influence of meteorological conditions on PM2.5 concentrations across China: A review of methodology and mechanism. *Environ. Int.* **2020**, *139*, 105558.

61. Tursumbayeva, M.; Kerimray, A.; Karaca, F.; Permadi, D.A. Planetary boundary layer and its relationship with PM2. 5 concentrations in Almaty 2022, Kazakhstan. *Aerosol Air Qual. Res.* **2022**, *22*, 210294.

62. Zender-Świercz, E.; Galiszewska, B.; Telejko, M.; Starzomska, M. The effect of temperature and humidity of air on the concentration of particulate matter—PM2.5 and PM10. *Atmos. Res.* **2024**, *312*, 107733.

63. Di, Q.; Kloog, I.; Koutrakis, P.; Lyapustin, A.; Wang, Y.; Schwartz, J. Assessing PM2.5 exposures with high spatiotemporal resolution across the continental United States. *Environ. Sci. Technol.* **2016**, *50*, 4712–4721.

64. Lyapustin, A.; Martonchik, J.; Wang, Y.; Laszlo, I.; Korkin, S. Multiangle implementation of atmospheric correction (MAIAC): 1. Radiative transfer basis and look-up tables. *J. Geophys. Res. Atmos.* **2011**, *116*, 3210.

65. Lyapustin, A.; Wang, Y.; Korkin, S.; Huang, D. MODIS Collection 6 MAIAC algorithm. *Atmos. Meas. Tech.* **2018**, *11*, 5741–5765.

66. Liang, F.; Xiao, Q.; Wang, Y.; Lyapustin, A.; Li, G.; Gu, D.; Pan, X.; Liu, Y. MAIAC-based long-term spatiotemporal trends of PM2.5 in Beijing, China. *Sci. Total Environ.* **2018**, *616–617*, 1589–1598.

67. Zhang, Z.; Wu, W.; Fan, M.; Wei, J.; Tan, Y.; Wang, Q. Evaluation of MAIAC aerosol retrievals over China. *Atmos. Environ.* **2019**, *202*, 8–16.

68. LAADS DAAC. NASA. 2024. Available online: **https://ladsweb.modaps.eosdis.nasa.gov/** (accessed on 16 January 2025).

69. Dee, D.P.; Uppala, S.M.; Simmons, A.J.; Berrisford, P.; Poli, P.; Kobayashi, S.; Andrae, U.; Balmaseda, M.A.; Balsamo, G.; Bauer, P.; et al. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **2011**, *137*, 553–597.

70. Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **2020**, *146*, 1999–2049.

71. GMTED2010. USGS. 2024. Available online: **https://www.usgs.gov/coastal-changes-and-impacts/gmted2010** (accessed on 16 January 2025).

72. Kamalov, F.; Atiya, A.F.; Elreedy, D. Partial Resampling of Imbalanced Data. *arXiv* **2020**, arXiv:2207.04631.

73. Cui, B.; Liu, M.; Li, S.; Jin, Z.; Zeng, Y.; Lin, X. Deep Learning Methods for Atmospheric PM2.5 Prediction: A Comparative Study of Transformer and CNN-LSTM-Attention. *Atmospheric Pollution Research* **2023**, *14*, 101833.

74. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.

75. Al-qaness, M.A.A.; Dahou, A.; Ewees, A.A.; Abualigah, L.; Huai, J.; Abd Elaziz, M.; Helmi, A.M. ResInformer: Residual Transformer-Based Artificial Time-Series Forecasting Model for PM2.5 Concentration in Three Major Chinese Cities. *Mathematics* **2023**, *11*, 476.

76. Dai, Z.; Ren, G.; Jin, Y.; Zhang, J. Research on PM2.5 Concentration Prediction Based on Transformer. *J. Phys. Conf. Ser.* **2024**, *2813*, 012023.

77. Srirenganathan Malarvizhi, A.; Pan, P. Multi-source data fusion for filling gaps in satellite Aerosol Optical Depth (AOD) using generative models. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Spatial Big Data and AI for Industrial Applications (GeoIndustry), Atlanta, GA, USA, 29 October 2024; pp. 28–38.

78. Malarvizhi, A.S.; Liu, Q.; Trefonides, T.S.; Hasheminassab, S.; Smith, J.; Huang, T.; Yang, C. The spatial dynamics of Ukraine air quality impacted by the war and pandemic. *Int. J. Digit. Earth* **2023**, *16*, 3680–3705.

Code: https://github.com/phobopan/dataaugmentation