atmosphere

MDPI

Article

Data Augmentation Strategies for Improved PM2.5 Forecasting Using Transformer Architectures

Phoebe Pan [1], Anusha Srirenganathan Malarvizhi [2] and Chaowei Yang [2,*]

[1] Thomas Jefferson High School for Science and Technology, Alexandria, VA 22312, USA; 2025ppan@tjhsst.edu
[2] Geography & Geoinformation Science Department, George Mason University, Fairfax, VA 22030, USA; asrireng@gmu.edu
* Correspondence: cyang3@gmu.edu

Abstract: Breathing in fine particulate matter of diameter less than 2.5 μm ($PM_{2.5}$) greatly increases an individual's risk of cardiovascular and respiratory diseases. As climate change progresses, extreme weather events, including wildfires, are expected to increase, exacerbating air pollution. However, models often struggle to capture extreme pollution events due to the rarity of high $PM_{2.5}$ levels in training datasets. To address this, we implemented cluster-based undersampling and trained Transformer models to improve extreme event prediction using various cutoff thresholds (12.1 μg/m^3 and 35.5 μg/m^3) and partial sampling ratios (10/90, 20/80, 30/70, 40/60, 50/50). Our results demonstrate that the 35.5 μg/m^3 threshold, paired with a 20/80 partial sampling ratio, achieved the best performance, with an RMSE of 2.080, MAE of 1.386, and R^2 of 0.914, particularly excelling in forecasting high $PM_{2.5}$ events. Overall, models trained on augmented data significantly outperformed those trained on original data, highlighting the importance of resampling techniques in improving air quality forecasting accuracy, especially for high-pollution scenarios. These findings provide critical insights into optimizing air quality forecasting models, enabling more reliable predictions of extreme pollution events. By advancing the ability to forecast high $PM_{2.5}$ levels, this study contributes to the development of more informed public health and environmental policies to mitigate the impacts of air pollution, and advanced the technology for building better air quality digital twins.

Keywords: air quality; $PM_{2.5}$ forecasting; data augmentation; cluster-based undersampling; Transformer model; digital twin; 2023 Canadian wildfires

Academic Editor: Xinghua Li

Received: 10 November 2024
Revised: 16 January 2025
Accepted: 17 January 2025
Published: 24 January 2025

Citation: Pan, P.; Malarvizhi, A.S.; Yang, C. Data Augmentation Strategies for Improved PM2.5 Forecasting Using Transformer Architectures. *Atmosphere* **2025**, *16*, 127. https://doi.org/10.3390/atmos16020127

1. Introduction

Air pollution remains one of the most pressing global health challenges, identified as the second leading risk factor for premature death worldwide. In 2021 alone, air pollution was responsible for approximately 8.1 million deaths globally, underscoring its profound impact on human health [1]. Fine particulate matter refers to particles with an aerodynamic diameter of 2.5 μm or less ($PM_{2.5}$). $PM_{2.5}$ particles are particularly concerning because they are small enough to penetrate deep into the lungs and even enter the bloodstream, posing significant risks to human health. The Global Burden of Disease (GBD) study estimated that ambient $PM_{2.5}$ exposure was responsible for approximately 4.14 million deaths globally in 2019 [2]. These particles are associated with a wide range of health outcomes, including stroke, ischemic heart disease, chronic obstructive pulmonary disease (COPD), and lung cancer [3–8]. The respiratory system, especially the lungs, is vulnerable to $PM_{2.5}$-induced toxicity, leading to inflammation and impaired immune responses, increasing susceptibility to respiratory infections [9]. Growing evidence suggests that $PM_{2.5}$ exposure is also linked

to neurodegenerative diseases. The small size of the particles enables them to penetrate the brain via the olfactory nerve [10]. Recent trends have shown an alarming increase in $PM_{2.5}$ emissions due to wildfires, exacerbated by climate change and land management practices. Wildfire-related $PM_{2.5}$ pollution has been observed to travel long distances, affecting regions far beyond the initial fire location [11]. Wildfires in the western United States have increased in frequency and intensity since the mid-1980s, primarily driven by rising temperatures and earlier spring snowmelt [12]. Climate projections suggest that the area affected by wildfires in the western U.S. could expand by 54% between 2046 and 2055 compared to 1996–2005 [13]. During severe wildfire events, $PM_{2.5}$ levels can spike to hazardous levels, exceeding the Environmental Protection Agency's (EPA) threshold of 225.5 $\mu g/m^3$ for hazardous air quality [14]. Given the severe health impacts and increasing frequency of extreme $PM_{2.5}$ pollution events, it is critical to implement proactive measures such as improved air quality monitoring, stricter emission control policies, and enhanced public health advisories.

$PM_{2.5}$ forecasting is critical for protecting public health by enabling timely interventions, reducing exposure to hazardous air pollution, and supporting broader pollution management efforts. However, $PM_{2.5}$ forecasting is challenging due to the complex interactions between atmospheric chemistry, meteorological variability, and human activities, which cause rapid fluctuations in pollutant levels [15,16]. Additionally, capturing temporal variations and spatial distributions is crucial for effective exposure assessment and health impact evaluation. Various modeling approaches have been used for $PM_{2.5}$ forecasting, ranging from traditional statistical methods, such as Autoregressive Integrated Moving Average (ARIMA), to Artificial Intelligence and Machine Learning (AI/ML) models. These AI/ML techniques include nonlinear models like Support Vector Regression (SVR) and Artificial Neural Networks (ANNs), which have shown promise in capturing complex relationships in air quality data [17,18]. While ANNs have been widely applied for $PM_{2.5}$ forecasting, their shallow structures often limit feature learning in complex datasets [19]. Recent Deep Learning (DL)-based approaches, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) models, have significantly improved both spatial and temporal pattern modeling [20]. Hybrid models combining CNNs and LSTMs have further enhanced forecasting accuracy, particularly for datasets with spatiotemporal complexity [21,22]. However, DL models still face challenges such as vanishing gradients and limited long-term dependency modeling, as noted by [21]. Transformer models, initially designed for Natural Language Processing (NLP) [23], have shown promise for long-term $PM_{2.5}$ forecasting due to their ability to capture long-range dependencies [24]. Unlike recurrent models, Transformers rely on self-attention mechanisms that allow more efficient information flow across sequences [23]. In [25] introduced the Informer model, which improves temporal embeddings to learn non-stationary and long-range temporal dependencies. However, it focuses solely on "temporal attention" and overlooks spatial relationships between variables. In [26] tackled this by developing a graph Transformer that captures dynamic spatial dependencies, using sparse attention to trim less relevant nodes. In [27] further advanced this with the Spacetimeformer, which flattens multivariate time series to handle spatial and temporal influences. Recent models like the Sparse Attention-based Transformer (STN) by [28] effectively reduce time complexity while capturing long-term dependencies in $PM_{2.5}$ data. Similarly, [29] proposed the SpatioTemporal (ST)-Transformer, designed to improve spatiotemporal predictions of $PM_{2.5}$ concentrations in wildfire-prone areas.

An often-overlooked challenge in $PM_{2.5}$ forecasting is data imbalance, particularly in predicting high pollution levels. AI/ML models have demonstrated strong performance in forecasting $PM_{2.5}$ under lower concentrations but often struggle to accurately capture

extreme pollution events where $PM_{2.5}$ levels exceed 60 $\mu g/m^3$ [30]. Studies have consistently shown that $PM_{2.5}$ concentrations tend to be underestimated during severe pollution episodes, as high-value events are underrepresented in the training data [31,32]. This imbalance results from the rarity of extreme pollution spikes, making it difficult for models to generalize and predict these critical conditions effectively [33–35]. Although this challenge is well-known, relatively few studies have focused on solutions for improving predictions of extreme $PM_{2.5}$ levels [36–38]. An effective strategy to address this imbalance is data augmentation, which expands the training dataset by introducing varied and informative samples, improving data diversity and quality. This approach enhances the representation of underrepresented patterns, leading to better model robustness and generalization [39–41]. Undersampling and oversampling are data augmentation techniques developed to address the challenges of imbalanced datasets, each employing distinct strategies to adjust data distribution and improve model performance. Oversampling techniques aim to increase the representation of the minority class by generating or duplicating data points to improve data diversity and representation. Random oversampling, a simpler approach, duplicates minority class instances but can lead to overfitting issues in conventional models [42]. Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) are widely used oversampling methods to mitigate the effects of imbalanced datasets [43], with variants like SMOTE with k-means also being prominent ([44,45]. Conversely, undersampling techniques reduce the dominance of the majority class by removing data points, aiming to create a more balanced representation. Random undersampling deletes the majority of class instances randomly but risks information loss [42]. Undersampling methods are combined with clustering approaches to balance datasets while preserving data structure. This involves clustering the data into several clusters using methods such as k-means clustering and then selecting representative points from each cluster to minimize information loss [46,47]. Several studies have applied oversampling and undersampling techniques to address data imbalance and improve model performance in the context of $PM_{2.5}$ modeling. In [48] aimed to improve the estimation accuracy of high $PM_{2.5}$ concentrations by using an AugResNet model with random oversampling and SMOTE. While their approach improved performance on high-value $PM_{2.5}$ datasets, a limitation of the study was its focus on a single cutoff threshold and $PM_{2.5}$ retrieval rather than forecasting, which limits its broader applicability. In [49] employed LSTM, GRU, and hybrid GRU + LSTM models with linear interpolation for data augmentation, expanding the dataset without addressing the imbalance between high and low $PM_{2.5}$ concentrations. Their approach did not specifically target data imbalance, focusing on general dataset expansion, which can lead to overfitting, as synthetically increasing the dataset size does not introduce new variability. In [50] tackled the dataset shift problem between urban and rural $PM_{2.5}$ data, addressing differences in predictor variable density using multiple imputations by chained equations; however, this study focused on correcting biases caused by variable density disparities rather than general $PM_{2.5}$ forecasting, which limits its relevance to broader $PM_{2.5}$ prediction challenges.

The current research on $PM_{2.5}$ forecasting reveals critical gaps that need further investigation.

1. One major challenge is data imbalance, where high $PM_{2.5}$ concentration events, particularly during extreme pollution episodes such as wildfires, are significantly underrepresented in datasets. This imbalance often leads to poor model performance in forecasting these critical pollution levels, as models struggle to generalize effectively under such conditions [31–33].
2. Another gap is the limited application of augmentation techniques tailored to address this imbalance. While methods like SMOTE and ADASYN have shown promise in balancing datasets [48,49], their use in $PM_{2.5}$ forecasting, particularly for extreme

pollution events, remains limited. Most studies have focused on general dataset expansion without targeting rare, high-concentration events, which can result in overfitting rather than improved generalization [42].

3. Lastly, the underexplored potential of Transformer models presents another critical gap. Despite their success in long-term sequence modeling across various domains [22–24], Transformers have been insufficiently investigated for $PM_{2.5}$ forecasting, especially in urban environments where pollution poses significant health risks. Their ability to capture long-term dependencies and complex spatiotemporal patterns remains underutilized in extreme $PM_{2.5}$ event forecasting [25,29].

This study addresses the identified research gaps by applying data augmentation techniques, specifically cluster-based undersampling with varying majority-to-minority class ratios, to improve the representation of high $PM_{2.5}$ concentrations in the training data. The majority-minority cutoff thresholds are selected based on two EPA-defined criteria, emphasizing the importance of robust models capable of accurately forecasting elevated $PM_{2.5}$ levels in real-world scenarios. The study leverages a Transformer-based architecture with multi-head sparse attention to tackle the challenge of long-term dependency modeling inspired by models like Informer and Spacetimeformer. The specific research objectives are listed below:

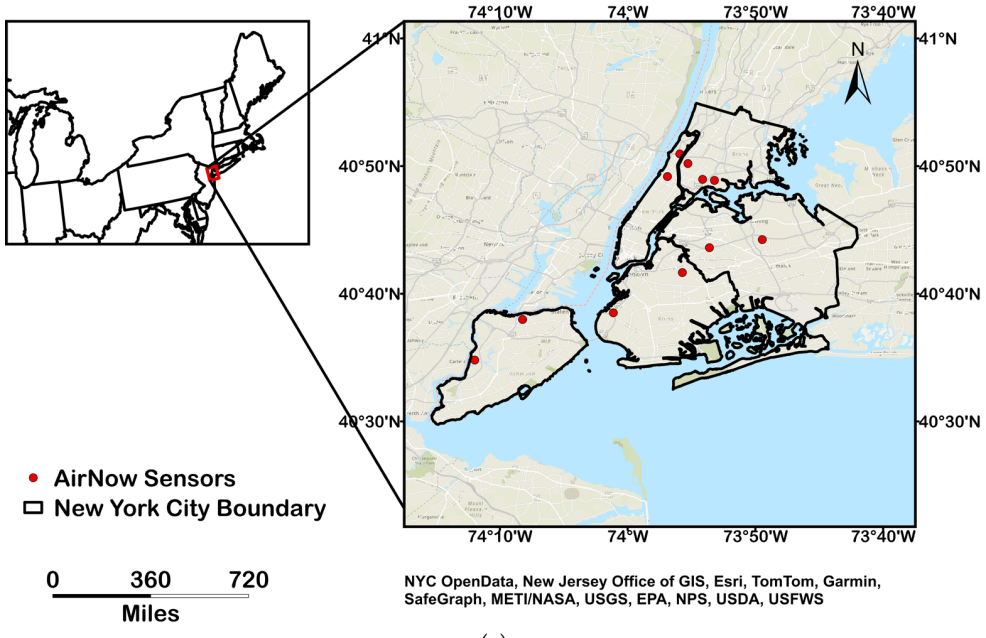
1. Augment imbalanced $PM_{2.5}$ dataset with cluster-based undersampling with different combinations of majority-to-minority class ratios.
2. Investigate the impact of two minority–majority cutoff thresholds based on limits set by the EPA on model performance.
3. Build and train a Transformer model to leverage the capabilities of multi-head attention in the context of $PM_{2.5}$ forecasting.
4. Develop a robust forecasting model that accurately predicts $PM_{2.5}$ concentrations, particularly during extreme pollution spikes caused by events like wildfires in New York City, Philadelphia, and Washington, D.C.

The remainder of this paper is organized as follows: Section 2 details the data, including the study area and data description. Section 3 outlines the methodology, covering data preprocessing, collocation, cutoff thresholds, cluster-based undersampling, the Transformer model architecture, and the model training and evaluation process. Section 4 presents experimental results, including accuracy assessment, partial sampling ratios, cutoff thresholds, and time series analysis. Section 5 discusses the findings, while Section 6 concludes the study with a summary of key insights and potential directions for future research.

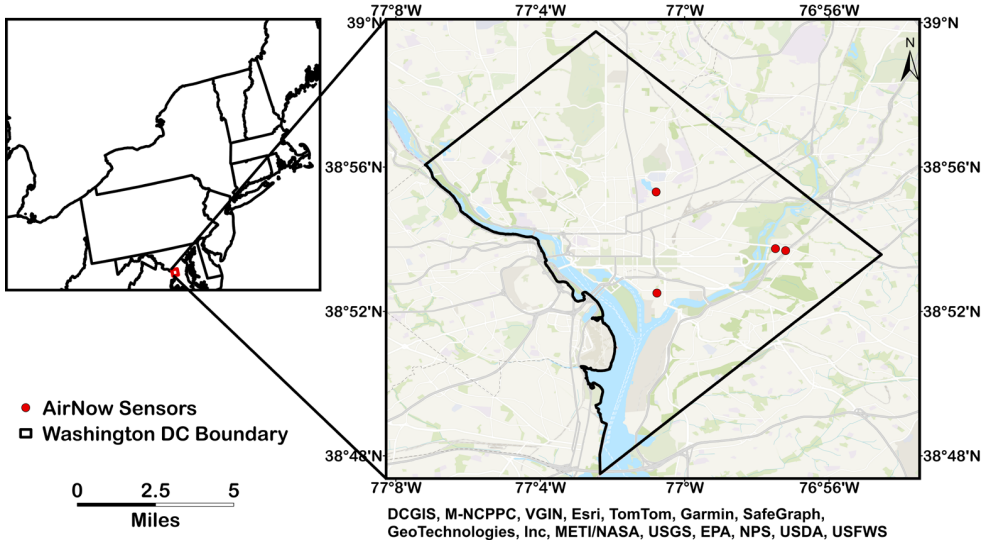
2. Data

2.1. Study Area

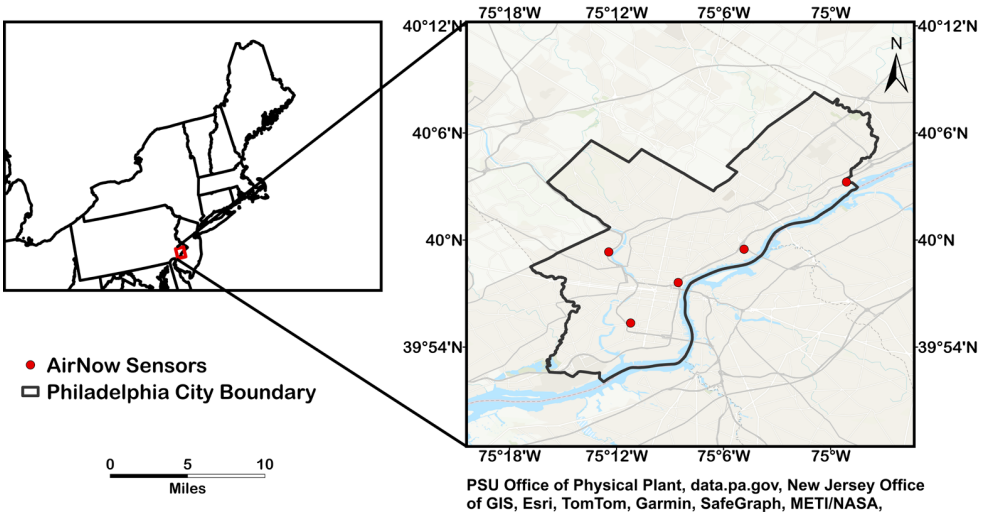
The study focuses on major urban areas in the northeastern United States, specifically New York City, Philadelphia, and Washington, D.C. Figure 1 depicts the locations of AirNow sensors in three areas: New York City (11 stations), Philadelphia (five stations), and Washington, D.C. (four stations). The area, population, and geographical locations of these three cities are listed in Table 1. These cities are characterized by high population densities, significant traffic volumes, and industrial activities, all of which contribute to elevated levels of air pollution. Urban areas are often hotspots for $PM_{2.5}$ due to vehicle emissions, industrial processes, and residential heating, making them critical regions for air quality monitoring and forecasting [51,52]. These urban environments also present challenges for air quality forecasting due to the complex interplay between local emissions and regional atmospheric transport processes.



(**a**)



(**b**)



(**c**)

Figure 1. Study area with AirNow sensor locations. (**a**) New York City, (**b**) Philadelphia, and (**c**) Washington D.C.

Table 1. Area, population, and geographical location of the three cities under consideration.

City	Area	Population	Coordinates
New York City	790 square km (302.6 square miles)	8.336 million	40.4774° N, −74.2591° W (southwest) to 40.9176° N, −73.7004° W (northeast)
Philadelphia	347.52 square km (134.18 square miles)	1.567 million	39.8670° N, −75.2803° W (southwest) to 40.1379° N, −74.9558° W (northeast)
Washington D.C.	76 square km (68 square miles)	671,803	38.7916° N, −77.1198° W (southwest) to 38.9955° N, −76.9094° W (northeast)

A significant event that affected air quality in 2023 was the Canadian wildfires, which profoundly impacted pollution levels across North America, particularly in urban areas of the northeastern United States [53,54]. The wildfires, which burned large swathes of forested areas in Canada, generated vast amounts of smoke and particulate matter that were transported southward by atmospheric winds, leading to unprecedented spikes in $PM_{2.5}$ concentrations in cities like New York, Philadelphia, and Washington, D.C. [55]. During this event, air quality in these cities reached hazardous levels, reducing visibility severely and prompting public health warnings [56]. $PM_{2.5}$ forecasting is critical in these urban areas because it helps predict and mitigate the health risks associated with high pollution levels, especially during extreme events like the 2023 Canadian wildfires. Accurate forecasting allows for timely public health warnings [57].

2.2. Data Description

Table 2 outlines the variables for forecasting $PM_{2.5}$ concentrations, with $PM_{2.5}$ from AirNow serving as the target variable. The covariates include aerosol optical depth (AOD) from the Moderate Resolution Imaging Spectroradiometer (MODIS) Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm. The AOD is a proxy for atmospheric particulate load linked to surface $PM_{2.5}$ concentration [58,59]. Several studies have demonstrated the strong relationship between meteorological factors and variations in $PM_{2.5}$ concentrations [58,60]. The meteorological variables include boundary layer height, relative humidity, temperature at 2 m, surface pressure, and speed, all sourced from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 dataset. These meteorological variables influence $PM_{2.5}$ size, concentration, dispersion, chemical transformation, and the formation of secondary particles. Boundary layer height influences vertical mixing and pollutant dispersion [61]. Relative humidity affects particle hygroscopic growth, while temperature can drive secondary aerosol formation through chemical reactions [62]. Surface pressure influences pollutant dispersion by controlling vertical air movement, and wind speed influences pollutant transport and dilution [60]. Additionally, elevation data sourced from the United States Geological Survey (USGS) was used as a geographical covariate. Elevation influences $PM_{2.5}$ by affecting air mixing and pollutant dispersion, with higher elevations typically experiencing stronger winds and lower atmospheric pressure, leading to reduced pollutant accumulation [63].

Table 2. Sources and units of variables used to forecast $PM_{2.5}$.

Variable	Type	Source	Unit
$PM_{2.5}$	Target	AirNow	$\mu g/m^3$
AOD	Covariate	MODIS MAIAC (Terra and Aqua)	Unitless
Boundary Layer Height	Covariate	ECMWF ERA5-hourly	Meter
Relative Humidity	Covariate	ECMWF ERA5-hourly	%
Temperature (at 2m)	Covariate	ECMWF ERA5-hourly	K
Surface Pressure	Covariate	ECMWF ERA5-hourly	Pa
Wind Speed	Covariate	ECMWF ERA5-hourly	m/s
Elevation	Covariate	USGS	Meter

2.2.1. Ground-Level $PM_{2.5}$ Measurements

Ground-level hourly $PM_{2.5}$ measurements were obtained from the EPA's AirNow program, which provides near-real-time air quality data, including $PM_{2.5}$ concentrations. The AirNow data undergo a rigorous quality control process before becoming publicly available. We downloaded $PM_{2.5}$ data from AirNow sensors in three major cities, New York City, Philadelphia, and Washington, D.C., through the AirNow API (http://airnowapi.org, accessed on 23 January 2025). These data are contributed by over 120 local, state, tribal, provincial, and federal government agencies participating in the AirNow program. It operates as a collaborative effort among multiple federal, state, and local agencies, ensuring quality control and consistency in air quality reporting. Given the need for timely analysis, we prioritized using near-real-time data over the delayed Air Quality System (AQS) data. These measurements were used to evaluate and compare model predictions of $PM_{2.5}$ concentrations in the selected urban areas.

2.2.2. Satellite-Derived AOD

The AOD data used in this study were derived from the MODIS aboard the Terra and Aqua satellites and processed using the MAIAC algorithm. Terra and Aqua provide daily AOD products at a spatial resolution of 1 km $\times$ 1 km, captured at approximately 10:30 and 13:30 local time, respectively [64,65]. MAIAC is an advanced algorithm designed for aerosol retrievals over dark vegetated surfaces and bright deserts, making it highly effective for air quality assessments due to its high spatial resolution [66,67]. The Version 6 MAIAC Land AOD product has been widely applied in air quality studies due to its superior spatial resolution and temporal coverage [65]. For this study, we used the MAIAC AOD product MCD19A2 at 550 nm and retained only high-quality AOD values, as indicated by the quality assessment flag marked "best quality". The data were sourced from the Level 1 and Atmosphere Archive and Distribution System Distributed Active Archive Center website [68].

2.2.3. Meteorological Variables

For this study, we utilized meteorological data from the ERA5 dataset, developed by the ECMWF. ERA5 is a comprehensive reanalysis product that estimates various atmospheric, land, and oceanic climate variables hourly. It offers continuous data from 1940 to the present with hourly temporal resolution, making it suitable for historical analysis and near-real-time applications. The dataset offers global coverage at a spatial resolution of $0.25^\circ \times 0.25^\circ$ on a regular latitude–longitude grid [69,70]. The data were sourced through the Copernicus Climate Data Store (C3S) in GRIB format, ensuring high detail and consistency for climate research. The meteorological variables selected for this study included boundary layer height (BLH), relative humidity, surface pressure (SP), 2 m temperature (T2M), and wind speed at 10 m (U10/V10). Wind speed was calculated from the eastward (U) and northward (V) wind components in mathematical terms as the square root of the sum of their squares.

2.2.4. Geographical Variables

This study used elevation data from the Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010) [71], a globally available dataset providing multi-scale topographic data. The GMTED2010 dataset offers three spatial resolutions: 30 arc-s (~1 km), 15 arc-s (~500 m), and 7.5 arc-s (~250 m) and elevation data are provided in raster format. It is generated by aggregating multiple global elevation sources using systematic subsampling to ensure comprehensive terrain representation. For this study, we used the 30 arc-s (~1 km) resolution data obtained from the USGS website [71].

3. Methodology

Figure 2 illustrates the workflow for the methodology used in this study to forecast $PM_{2.5}$ levels. Starting with data acquisition, the data undergo processing before selection of cutoff thresholds. Two cutoff thresholds are employed: 12.1 $\mu g/m^3$ and 35.5 $\mu g/m^3$, followed by applying various partial sampling ratios ranging from 10/90 to 50/50. Each threshold and sampling ratio combination is fed into a Transformer model for model training. The accuracy of each model is assessed, and the best-performing model is selected to produce $PM_{2.5}$ forecasts. This process allows for identifying the most effective threshold and sampling ratio, optimizing model performance for $PM_{2.5}$ prediction.

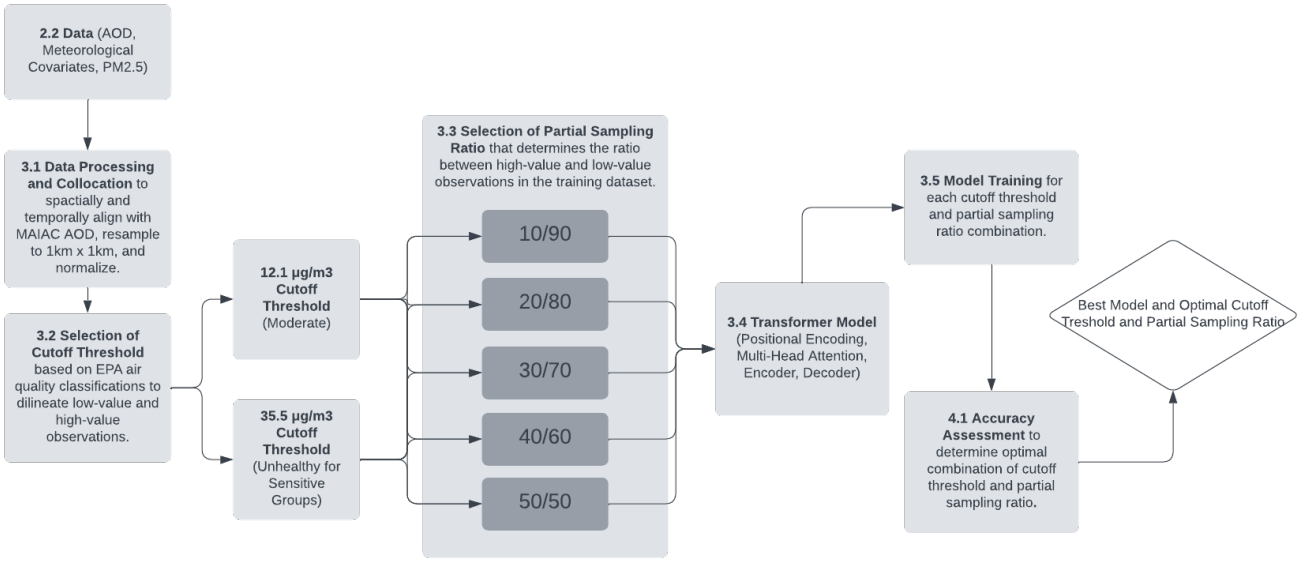


Figure 2. Methodology workflow diagram.

3.1. Data Preprocessing and Collocation

The spatiotemporal collocation was conducted by aligning the datasets with the MAIAC AOD data from MODIS to ensure consistency in time and space for seamless integration into the $PM_{2.5}$ forecasting model. This process involved matching all datasets to a common temporal observation window and a unified spatial resolution and projection. For temporal collocation, meteorological variables from the ERA5 ECMWF and the AirNow $PM_{2.5}$ datasets were matched to the Terra and Aqua MODIS satellite overpass times, approximately 10:30 and 13:30 local time. Daily averages of the hourly data were calculated within these overpass windows to ensure temporal consistency.

For spatial matching, the meteorological variables from ERA5 ECMWF, originally at a spatial resolution of $0.25^\circ \times 0.25^\circ$, were resampled to 1 km $\times$ 1 km using spatial interpolation and reprojected to the USA Contiguous Lambert Conformal Conic projection. Similarly, the point-based AirNow $PM_{2.5}$ sensor data were interpolated using kriging to create a continuous gridded surface, resampled to a 1 km $\times$ 1 km resolution, and reprojected to the same projection. The spatial matching was performed using ArcGIS Pro 3.2.1.

Data normalization is a crucial preprocessing step for many machine learning models, especially in deep learning applications, as features with a wide range can cause instability during model training. Therefore, we used the min-max scaler method to linearly transform the raw data to a value between 0 and 1 to eliminate dimensional effects. The formula is specified in Equation (1).

$$x' = \frac{x - \mu}{\max(x) - \min(x)} \quad (1)$$

where x and $x^{'}$ refer to the values before and after normalization, and min(x) and max(x) refer to the minimum and maximum values before normalization, respectively.

3.2. Cutoff Threshold

In their study, [48] defined 75 $\mu g/m^3$ as the cutoff threshold to distinguish low-value from high-value $PM_{2.5}$ points, aligning with China's air quality standard, which classifies 75 $\mu g/m^3$ as the lower limit for light $PM_{2.5}$ pollution. The selection of a cutoff value for distinguishing between majority and minority classes plays a crucial role in determining class distribution and, consequently, the model's performance in forecasting high-pollution periods. In the context of air quality in the United States, we used $PM_{2.5}$ classifications by the EPA to determine cutoff threshold values [14]. To evaluate the sensitivity of model performance to different cutoff values, this study compares two thresholds aligned with EPA standards: 12.1 $\mu g/m^3$ (Moderate) and 35.5 $\mu g/m^3$ (Unhealthy for Sensitive Groups). This approach represents a novel contribution to the field, as no prior research has investigated the interaction between data augmentation and cutoff threshold in the context of $PM_{2.5}$ forecasting. Existing studies typically rely on a single threshold value without examining its suitability for capturing the dynamics of air quality in their specific geographic region [48].

The total dataset in this study consisted of 4,026,240 valid points. With a cutoff threshold of 12.1 $\mu g/m^3$, 3,472,689 points were classified as "low-value", and 553,551 points were classified as "high-value". The ratio of high-value points to low-value ones was approximately 4:25 in the whole dataset. Ratios by city and the specific number of low- and high-value points are listed in Table 3.

Table 3. Breakdown of low- and high-value points with 12.1 $\mu g/m^3$ cutoff threshold.

City	Total Points	Low-Value	High-Value	Ratio of High- to Low-Value
New York City	2,284,200	2,038,105	246,095	0.1207
Washington DC	456,840	406,716	50,124	0.1232
Philadelphia	1,285,200	1,027,868	257,332	0.2503
Total	4,026,240	3,472,689	553,551	0.1594

With a cutoff threshold of 35.5 $\mu g/m^3$, 4,004,433 points were classified as "low-value", and 21,807 points were classified as "high-value". The ratio of high-value points to low-value ones was approximately 5:1000 in the whole dataset. Ratios by city are listed in Table 4 alongside the specific number of low- and high-value points.

Table 4. Breakdown of low- and high-value points with 35.5 $\mu g/m^3$ cutoff threshold.

City	Total Points	Low-Value	High-Value	Ratio of High- to Low-Value
New York City	2,284,200	2,272,914	11,286	0.00496
Washington DC	456,840	454,649	2191	0.00481
Philadelphia	1,285,200	1,276,870	8330	0.00652
Total	4,026,240	4,004,433	21,807	0.00544

3.3. Cluster-Based Undersampling

In this study, we implemented cluster-based undersampling to address class imbalance in the training data. The first step involved grouping data points into clusters using the k-means algorithm, which organized the data based on feature similarities. This clustering approach preserved the inherent structure of the dataset by ensuring that similar data points were grouped together, which is crucial for maintaining data integrity when performing

undersampling. By applying the undersampling strategy within each cluster, we selected a subset of instances, effectively reducing the majority class without losing the diversity within the data. This method allowed for a more representative sample, ensuring that both majority and minority classes were evenly distributed. This represents a novel approach because, unlike past research that predominantly relied on oversampling techniques such as random oversampling or linear interpolation—methods prone to overfitting—we focused on reducing the majority class to address class imbalance [42]. By integrating clustering into the undersampling process, we ensured that the selected subset retained the diversity and representativeness of the original dataset, providing a balanced yet structurally faithful training sample [46].

Before applying data augmentation, 20% of the original dataset was set aside for testing. The remaining 80% was used for model training, with points selected randomly based on the data augmentation technique. This approach ensured that each model was trained on datasets with unique sampling strategies, but all models were evaluated against a consistent testing dataset. The testing dataset was intentionally designed to mirror the original data distribution, ensuring fair comparisons across models trained on different augmented datasets.

Many studies aim to achieve a perfect 50/50 balance between minority and majority class points. However, this idealized ratio is not always the most effective for model training, especially when dealing with environmental data like $PM_{2.5}$, where the natural distribution is often skewed. Partial sampling, as discussed by [72], involves adjusting the class ratio to values between the original class distribution and an equal 50/50 split. This technique provides a more nuanced approach, reflecting real-world distributions more accurately while improving model generalizability.

In [73] found that a minority-to-majority class ratio of approximately 0.75 was optimal in many scenarios, based on a systematic review of datasets and resampling techniques. Following this insight, the present study applied various partial sampling ratios to explore whether these findings hold for $PM_{2.5}$ forecasting models. The aim was to determine whether a similar class balance could yield improved prediction accuracy in this context.

Initially, the training dataset contained around 3 million points. However, after data augmentation and resampling, each training dataset was reduced to approximately 35,000 points. Although all training datasets had the same number of observations, the distribution of $PM_{2.5}$ values varied according to the selected threshold and resampling ratio, reflecting the impact of these parameters on the dataset's composition. Table 5 shows the exact number of high-value and low-value points for each partial sampling ratio dataset. Although the two cutoff thresholds result in the same number of high-value and low-value points at each partial sampling ratio, their classification of high and low values differ, leading to distinct distributions across the datasets.

Table 5. Number of high- and low-value points in the training dataset at each partial sampling ratio.

Partial Sampling Ratio	High-Value Points	Low-Value Points
10/90	3498	31,482
20/80	6996	27,984
30/70	10,494	24,486
40/60	13,992	20,988
50/50	17,490	17,490

Figure 3 presents the distributions of training datasets across partial sampling ratios using a 12.1 $\mu g/m^3$ threshold, with different partial sampling ratios of 10/90, 20/80, 30/70, 40/60, and 50/50. These sampling ratios represent the proportion of minority (high $PM_{2.5}$) to majority (low $PM_{2.5}$) points included in the dataset.

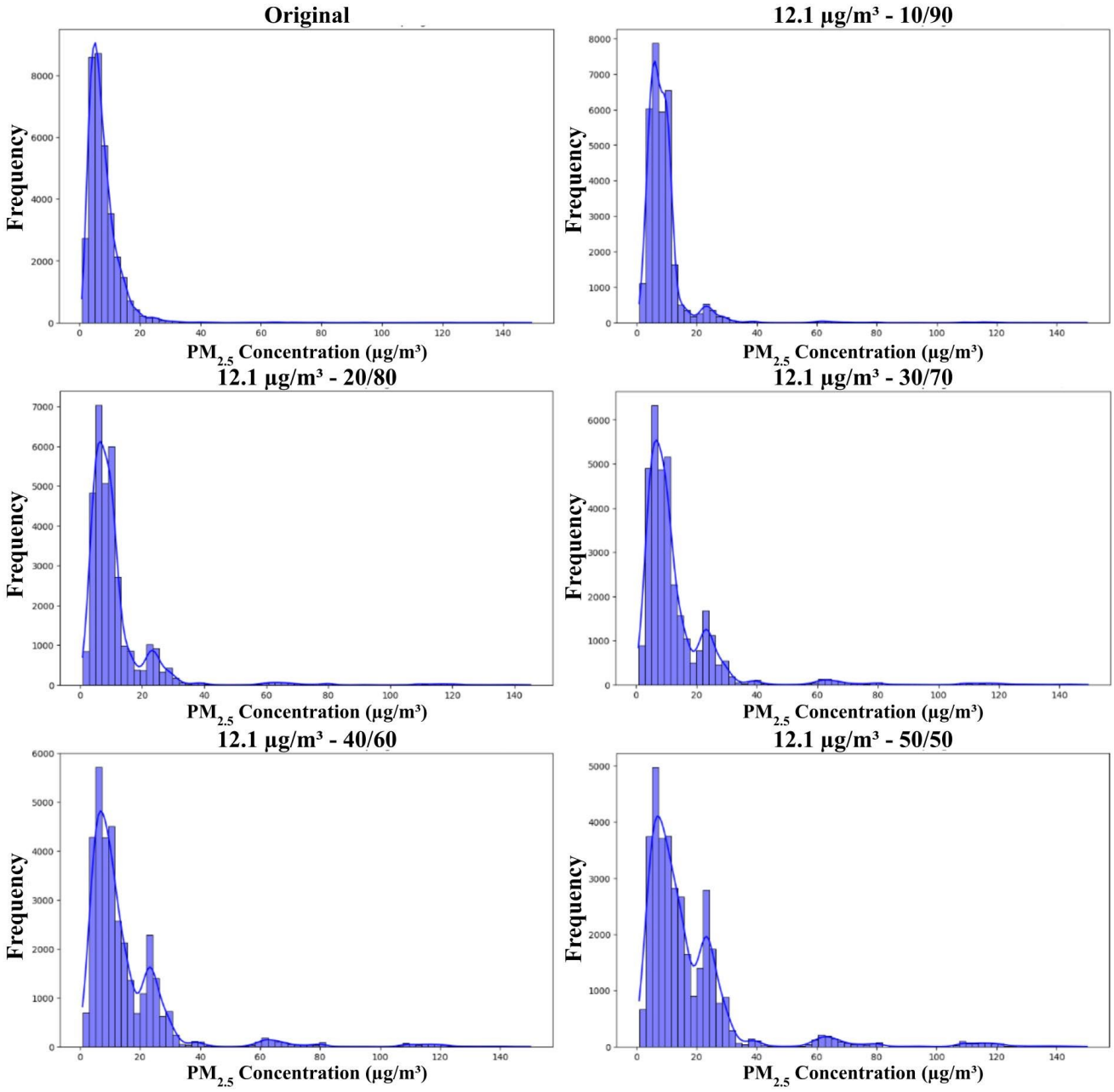


Figure 3. Distribution of training dataset with cutoff threshold 12.1 µg/m^3 and partial sampling ratio, from top left to bottom right, of none (original distribution), 10/90, 20/80, 30/70, 40/60, 50/50.

The graphs in Figure 4 present the distributions of training datasets across partial sampling ratios using a 35.5 µg/m^3 threshold, with different partial sampling ratios of 10/90, 20/80, 30/70, 40/60, and 50/50.

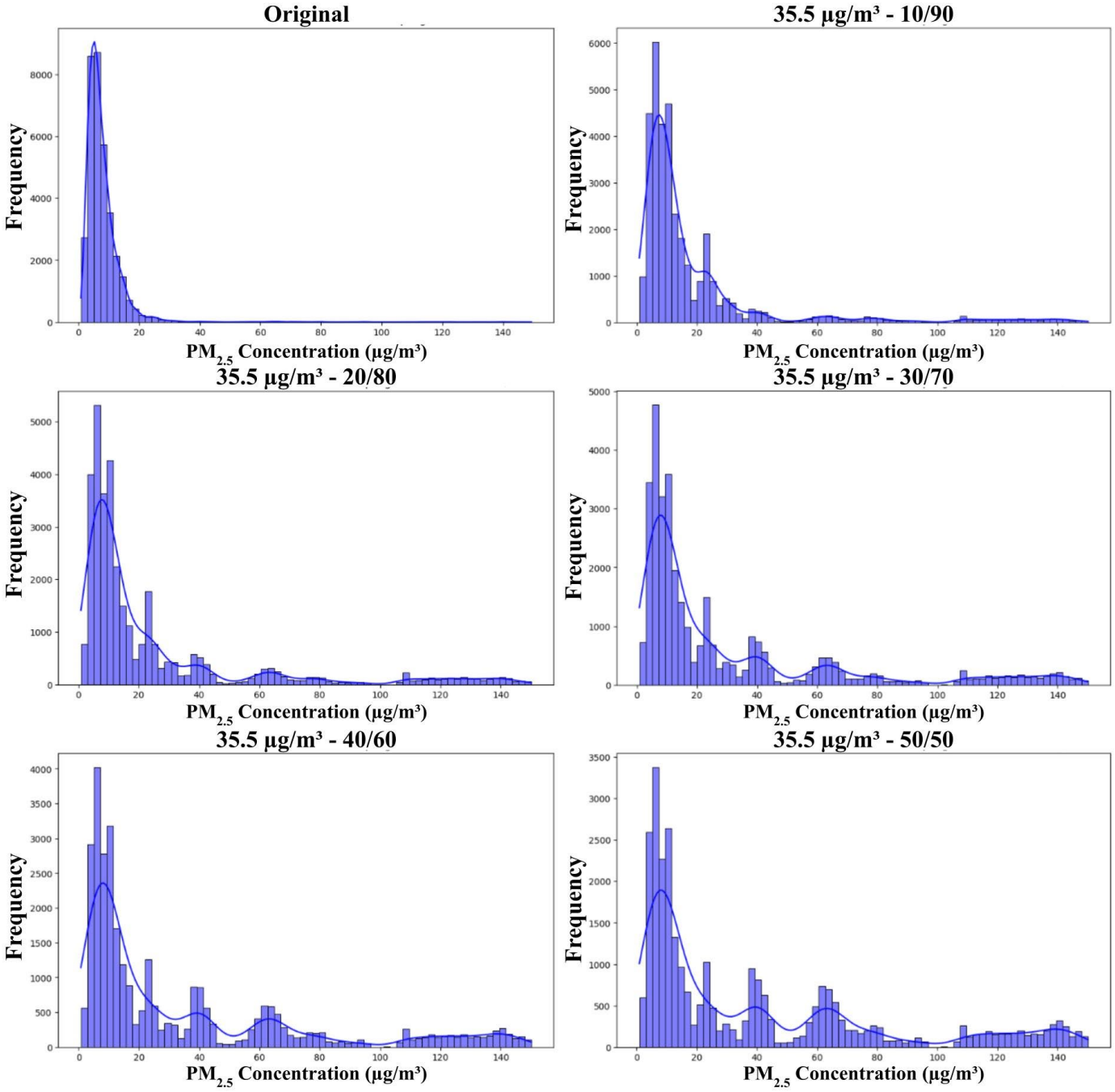


Figure 4. Distribution of training dataset with cutoff threshold 35.5 µg/m^3 and partial sampling ratio, from top left to bottom right, of none (original distribution), 10/90, 20/80, 30/70, 40/60, 50/50.

3.4. Transformer Model Architecture

The Transformer model has revolutionized various domains of ML, including NLP and time series forecasting [23]. In the context of $PM_{2.5}$ forecasting, the Transformer model's ability to capture long-range dependencies and complex temporal patterns makes it a powerful tool for forecasting air pollution levels [28,73]. Traditional methods often struggle with the non-linear and dynamic nature of $PM_{2.5}$ data, but the Transformer's self-attention mechanism allows it to weigh the importance of different time steps effectively, leading to more accurate and robust forecasts. While the Transformer model has shown great promise in $PM_{2.5}$ forecasting, this study is the first to explore the impact of data augmentation techniques on Transformer-based models in this domain, introducing a novel approach

to enhance performance and robustness. A simplified diagram of the architecture of the Transformer model is displayed in Figure 5.

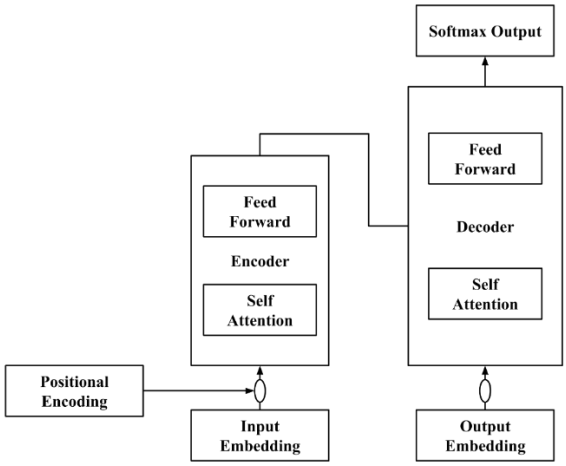


Figure 5. Transformer model architecture.

3.4.1. Positional Encoding

A Transformer model differentiates itself from traditional convolutional and recurrent neural networks by employing a novel positional encoding mechanism to preserve temporal relationships. This is achieved by embedding sine and cosine functions of varying frequencies into the normalized input sequences as illustrated by Equations (2) and (3), respectively.

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{2}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{3}$$

Here, pos represents the position of a data point within the sliding window, and i indicates the i-th dimension in the feature space. This approach allows the Transformer to retain the order of the sequence data, ensuring that the temporal dynamics are preserved and effectively leveraged during training and inference [23].

3.4.2. Multi-Head Attention

To make the model focus on assigning different weights to the input time series information during the encoding phase, an attention mechanism is often used to quantify the dependencies between them. The attention score determines the extent to which the information corresponding to a time slice in the time series should be focused on in future forecasts, and it can be calculated using Equation (4).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)^V \tag{4}$$

where Q, K, and V are the matrices of the queries, keys, and values, respectively, and d_k is the dimension of the key (K).

The multi-head attention mechanism enhances the model's ability to capture long-range dependencies by allowing it to focus on both sequence positions and multiple heads simultaneously. Each pollution factor's position, feature, and value are treated as separate heads, with multiple matrices applied to repeat the self-attention process across parallel

layers. This approach enables the model to consider various relationships between pollution factors and meteorological conditions.

3.4.3. Encoder

In this study, the encoder consists of a stack of $n = 6$ identical layers. In each layer, the input goes through multi-head self-attention, where the same input is used for queries, keys, and values, and attention weights are computed based on the provided mask. The output from self-attention is added to the original input, normalized using LayerNorm, and passed through a feed-forward network. After the feed-forward computation, the result is again added to the input, followed by another layer normalization and dropout.

3.4.4. Decoder

Each decoder also consists of a stack of $n = 6$ layers. In each decoder layer, the first step applies self-attention, where the target sequence attends to itself, with a mask to control the attention. Next, cross-attention is applied, where the output from the self-attention step attends to the encoder output, allowing the decoder to incorporate information from the encoder while applying a source mask. Finally, the result passes through a feed-forward network, and after each attention and feed-forward step, residual connections, normalization, and dropout are applied to maintain stability.

3.5. Model Training and Evaluation

3.5.1. Model Training and Hyperparameter Tuning

The data splitting method allowed models to be trained on varied datasets while being tested on a consistent, representative test set for fair comparison. More detailed dataset creation procedures can be found in Section 3.3.

We opted not to perform extensive hyperparameter tuning given that the hyperparameters specified in our experiments, as outlined in Table 6, already yielded satisfactory results. Adam was chosen due to its widespread use and proven effectiveness in training deep learning models, as highlighted in previous studies [23,74]. The parameters used in this study were determined through trial and error, like the approach by [73]. It is also important to note that the original authors of the Transformer did not perform extensive hyperparameter tuning, and many subsequent studies employing Transformers have followed a similar approach due to the high computational expense of such tuning [23,29,73–76].

Table 6. Hyperparameters used for model training.

Training Parameter	Values
Model training data	2021, 2022, 2023
Data split	Training (80%) and testing (20%)
Optimizer	Adam
Learning Rate	0.001
Epochs	20
Number of encoder and decoder layers	6
Model Dimension	8
Batch Size	256
Input length	8
Output length	8
Dropout Rate	0.1

To maintain the validity of comparisons across different models and experiments, we kept the hyperparameters constant throughout all tests. This decision ensured that performance differences could be attributed to model adjustments rather than variations in tuning.

3.5.2. Accuracy Measures

This paper employs Root Mean Square Error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2) as metrics for assessing model accuracy. RMSE evaluates the extent to which the predicted value curve aligns with the observed value curve. MAE measures the average absolute difference between the predicted and actual values. R^2 indicates the proportion of the variance in the dependent variable (y) that can be explained by the independent variable (x). The respective formulas for these calculations are shown in Equations (5)–(7).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \tag{7}$$

where n refers to the number of data, y_i refers to the ith observed value, $\hat{y}_i$ refers to the i^{th} predicted value, and $\overline{y}_i$ refers to the average of all observed values.

4. Experiments and Results

4.1. Accuracy Assessment

The optimal partial sampling ratios were selected by comparing model performance metrics across augmented datasets and the original, unaugmented dataset at each cutoff threshold.

For experiments performed with a cutoff threshold of 12.1 μg/m^3, accuracy metrics are displayed in Table 7. As the resampling ratio becomes more balanced, ranging from 10/90 to 50/50, both RMSE and MAE metrics generally decrease, indicating improved model performance. The best overall performance is observed at the 50/50 ratio, where the RMSE reaches 2.757, the MAE is 1.044, and R^2 achieves a value of 0.850. This R^2 value suggests that the 50/50 ratio offers the strongest correlation between forecasted and true $PM_{2.5}$ values, making it the most effective configuration for balanced data.

Table 7. Accuracy measurements of models trained on data augmented with cutoff threshold 12.1 μg/m^3 and different partial sampling ratios tested on the whole and high-value testing dataset.

Resampling Ratio	Whole			High-Value		
	RMSE	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**
Original	3.174	0.661	0.801	32.013	26.705	0.036
10/90	3.217	0.726	0.796	29.366	20.284	0.188
20/80	3.090	1.145	0.812	25.948	19.044	0.366
30/70	2.823	1.535	0.843	25.243	18.827	0.400
40/60	2.816	1.325	0.845	23.284	17.383	0.490
50/50	2.757	1.044	0.850	21.287	14.114	0.574

Accuracy metrics for experiments performed with a cutoff threshold of 35.5 μg/m^3 are shown in Table 8. Interestingly, the 20/80 resampling ratio emerges as the optimal

configuration overall, achieving the lowest RMSE (2.080) and MAE (1.386), alongside the highest R^2 value of 0.914. This strong performance suggests that a 20/80 ratio balances the trade-off between capturing minority and majority points while minimizing error. The same ratio also delivers the best results for high-value $PM_{2.5}$ points, with an RMSE of 15.353, MAE of 10.077, and an R^2 value of 0.778, demonstrating that it is particularly effective for extreme pollution levels.

Table 8. Accuracy measurements of models trained on data augmented with cutoff threshold 35.5 μg/m^3 and different partial sampling ratios tested on the whole and high-value testing dataset.

Resampling Ratio	Whole			High-Value		
	RMSE	MAE	R^2	RMSE	MAE	R^2
Original	3.174	0.661	0.801	41.34	28.269	0.607
10/90	2.282	1.592	0.897	19.747	13.81	0.633
20/80	2.080	1.386	0.914	15.353	10.077	0.778
30/70	2.306	1.671	0.895	16.095	12.204	0.756
40/60	2.423	1.726	0.884	16.556	12.917	0.741
50/50	2.677	1.875	0.858	19.116	14.321	0.656

When comparing the performance of models trained on the original dataset to those trained on resampled datasets, the original data consistently underperforms, particularly in terms of error metrics like RMSE and R^2. This pattern emphasizes the value of resampling techniques for improving model accuracy.

4.2. Partial Sampling Ratio Comparison

At a cutoff threshold of 12.1 μg/m^3, in evaluating model performance across varying partial sampling ratios, tests on both the full dataset and high-value points demonstrate a clear trend: the 50/50 partial sampling ratio consistently yields optimal results, as displayed in Figure 6. For the full dataset, RMSE decreases as the sampling ratio becomes more balanced, reaching its lowest point at the 50/50 ratio. This indicates that more balanced data distribution significantly enhances forecast accuracy. Similarly, the R^2 value steadily increases, peaking at the 50/50 ratio, signaling the model's improved ability to capture long-range dependencies at this balanced ratio.

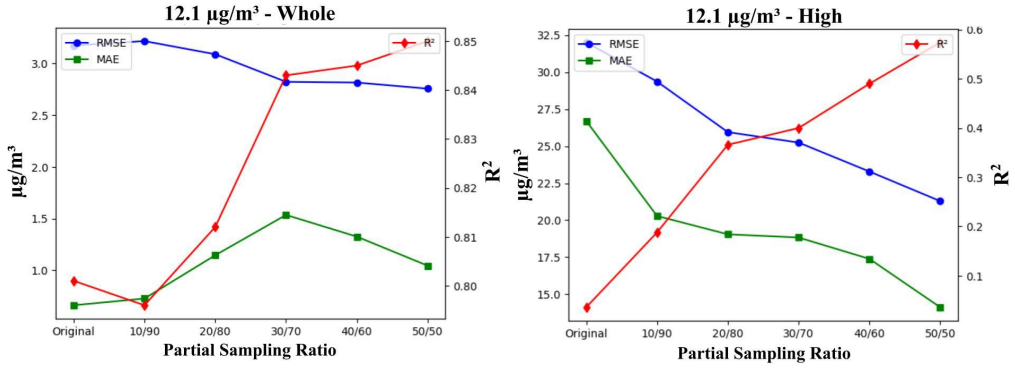


Figure 6. RMSE (blue), MAE (green), and R^2 (red) performance metrics across various partial sampling ratios using the 12.1 μg/m^3 threshold tested on whole (**left**) and high-value (**right**) testing datasets.

For high-value points, the results further underscore the importance of balanced resampling. RMSE shows a marked decline, and MAE gradually reduces as the ratio approaches 50/50. The model's highest R^2 value at this ratio confirms its strongest performance in predicting high-value points with greater accuracy. Overall, the 50/50 sampling ratio

emerges as the optimal configuration, demonstrating that more evenly distributed data enhances the model's performance, particularly in forecasting high-value events.

Patterns of model performance across varying partial sampling ratios change for the cutoff threshold of 35.5 μg/m^3, as presented in Figure 7. For the whole dataset, RMSE decreases as the partial sampling ratio becomes more balanced, reaching its minimum at 20/80. However, as the ratio becomes more balanced at 30/70, 40/60, and 50/50, RMSE slightly increases, indicating that the most balanced ratios do not necessarily lead to the best performance. Also, the R^2 value peaks at 20/80 but declines for more balanced ratios, suggesting that more even data distribution does not always improve model performance.

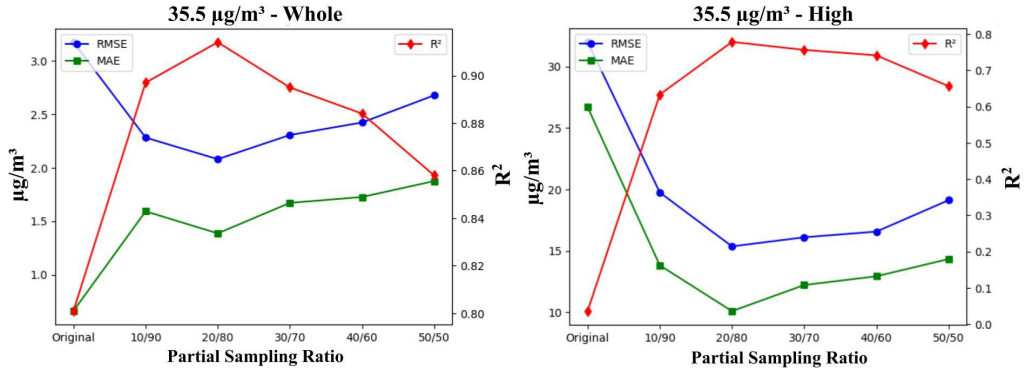


Figure 7. RMSE (blue), MAE (green), and R^2 (red) performance metrics across various partial sampling ratios using the 35.5 μg/m^3 threshold tested on whole (**left**) and high-value (**right**) testing datasets.

For high-value points, RMSE shows a sharp decline from its value based on the original data, continuing to decrease at the 20/80 ratio, with further stabilization beyond this point. MAE follows a similar trend, with a steep drop at 20/80 and stabilization thereafter. This indicates that the 20/80 partial sampling ratio effectively minimizes errors for high-value points. Similarly, R^2 improves significantly with slightly more balanced resampling, reaching its peak at 20/80, and begins to drop afterward, highlighting the model's best performance at this ratio.

Overall, the 20/80 ratio provides the best performance for both the full dataset and high-value points, delivering the lowest RMSE and highest R^2. Models trained on the original data perform the worst in terms of RMSE and R^2, underscoring the value of resampling for improving forecast accuracy, particularly for high-value points.

The discrepancy between RMSE and MAE in the original dataset arises from the nature of these metrics. RMSE amplifies the impact of large errors due to its squaring mechanism, making it highly sensitive to outliers, whereas MAE treats all errors equally, offering a more robust reflection of average performance (Chai and Draxler, 2014; Willmott and Matsuura, 2005). This suggests that the original dataset likely contains a few large outliers that inflate RMSE without significantly affecting MAE. As the partial sampling ratio becomes more balanced, the model improves accuracy when predicting high-value outliers (leading to lower RMSE) but loses some accuracy in predicting low-value events (causing a slight increase in MAE).

4.3. Cutoff Threshold Comparison

Models trained on the 35.5 μg/m^3 threshold consistently outperformed those trained on the 12.1 μg/m^3 threshold in terms of RMSE and R^2, as demonstrated in Figure 8. The resampling ratio plays a crucial role in model performance, with the 20/80 ratio emerging as optimal for the 35.5 μg/m^3 threshold, while the 50/50 ratio works best for the 12.1 μg/m^3 threshold. This disparity is largely driven by the nature of the data captured at each

threshold. The higher 35.5 μg/m^3 threshold likely includes a more concentrated set of high-value points, making a less balanced ratio like 20/80 more effective since the distinct minority points do not require as much balancing. In contrast, the 12.1 μg/m^3 threshold includes more low-value points, necessitating a 50/50 ratio to adequately represent both minority and majority groups.

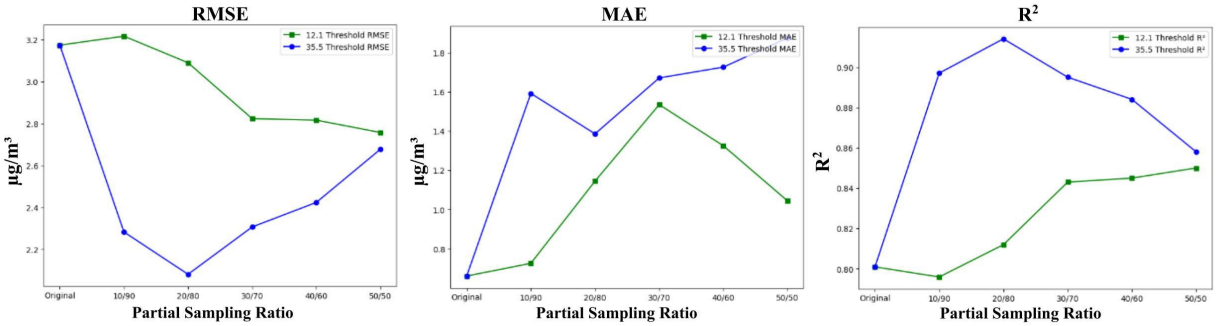


Figure 8. RMSE, MAE, and R^2 performance metrics for the 12.1 μg/m^3 (green) and 35.5 μg/m^3 (blue) thresholds across partial sampling ratios.

RMSE, which squares the differences before averaging, amplifies larger errors, making it more sensitive to a few large deviations from actual values. This explains why models trained on the 12.1 μg/m^3 threshold performed worse in terms of RMSE, as the larger prediction errors had a greater impact. However, MAE, which treats all errors equally, performed better for the 12.1 μg/m^3 threshold, indicating that while the errors were frequent, they were smaller in magnitude.

For the 12.1 μg/m^3 threshold, RMSE consistently decreases as the sampling ratio becomes more balanced, reaching its minimum at the 50/50 ratio. This trend highlights the importance of equal representation of minority and majority classes in improving overall accuracy for a lower threshold value. Conversely, MAE initially increases from 10/90 to 30/70 but significantly improves at 50/50. R^2 also shows steady improvement with increasing balance, reaching its peak at 50/50, where the model captures the strongest correlation between predicted and actual values.

For the 35.5 μg/m^3 threshold, RMSE shows a sharp decrease as the resampling ratio shifts from 10/90 to 20/80, indicating improved model performance by reducing large prediction errors. However, as the ratio becomes more balanced beyond 20/80, RMSE starts to increase slightly, suggesting that the model begins to overfit to the minority class while losing accuracy for low-value points. R^2 follows a similar trend, increasing to a peak at 20/80, where it achieves the highest variance explanation. Beyond this optimal ratio, R^2 declines, reflecting the reduced ability to accurately capture the distribution of both high- and low-value points.

The chosen thresholds, 12.1 μg/m^3 and 35.5 μg/m^3, while aligned with EPA air quality standards, have inherent limitations that may impact the generalizability of the findings. First, these thresholds are specific to U.S. regulatory definitions and may not capture the nuances of air quality classifications used in other regions, such as Europe or Asia, limiting the global applicability of the model. Additionally, these fixed thresholds may oversimplify the dynamic and continuous nature of $PM_{2.5}$ pollution levels, potentially misclassifying borderline cases and reducing sensitivity in capturing real-world fluctuations. The reliance on static thresholds also fails to account for seasonal or geographic variations in $PM_{2.5}$ concentrations, which could alter the distribution of high- and low-value points and affect model training. Furthermore, by focusing only on two thresholds, the study may overlook potential insights that could arise from exploring a wider range of cutoff values, especially for datasets with different pollutant distributions. These limitations highlight the need for

future work to explore adaptive or region-specific thresholds and assess their impact on model performance.

4.4. Time Series Analysis

Figure 9 presents the time series comparison of observed and forecasted $PM_{2.5}$ between the models trained on the original versus augmented dataset with cutoff threshold of 35.5 $\mu g/m^3$ and partial sampling ratio of 20/80. For all three cities, the model trained on the original dataset shows strong accuracy for lower $PM_{2.5}$ concentrations, particularly for values below 30 $\mu g/m^3$. This is reflected in the high similarity between forecasted and observed values at these low levels. However, the model struggles to predict higher $PM_{2.5}$ concentrations, reaching a ceiling in magnitude when faced with extreme pollution events, as evidenced in the red-boxed regions. This limitation arises from the imbalanced dataset, where the majority of points consist of lower values, leading the model to prioritize these over the rarer high-value points. As a result, the model is unable to fully capture extreme $PM_{2.5}$ events, which shows that forecast accuracy tends to decline as $PM_{2.5}$ levels increase.

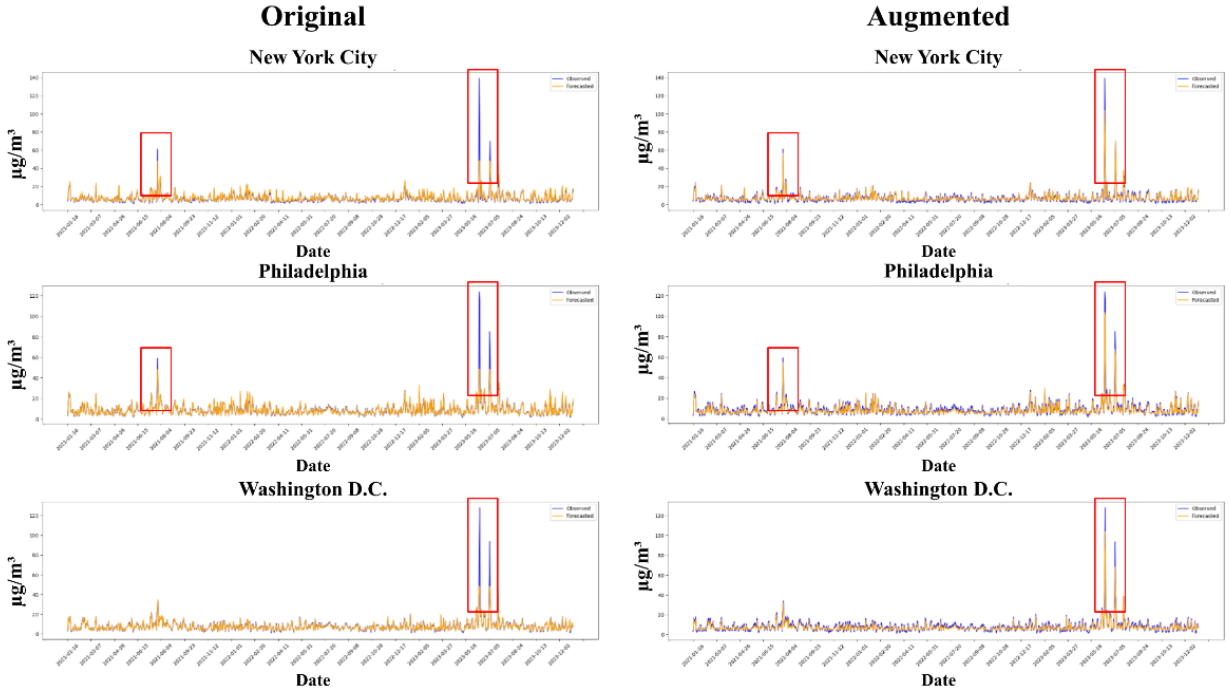


Figure 9. Time series of observed (blue) and forecast (orange) $PM_{2.5}$ concentrations from 2021 to 2023 in New York City (**top**), Philadelphia (**middle**), and Washington D.C. (**bottom**). The left column shows predictions from the model trained on original distribution and the right column shows 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.

In contrast, models trained on the augmented dataset, using a 35.5 $\mu g/m^3$ cutoff threshold and a 20/80 partial sampling ratio, demonstrate improved performance in capturing high-value $PM_{2.5}$ events. Although there is a trade-off, where the model's accuracy for lower $PM_{2.5}$ levels is slightly reduced, this adjustment leads to significantly better RMSE and R^2 measures. The forecasted values in the red-boxed regions are much closer to the observed peaks, demonstrating that the model trained on augmented data is better equipped to handle rare and extreme pollution levels. The trade-off is seen in the slightly worse MAE, as the augmented dataset introduces more diversity and some smaller errors that MAE treats equally, while RMSE emphasizes the larger improvements in extreme cases. The model built on the augmented data is better suited to handle high-value

points, which is particularly beneficial in scenarios where predicting extreme pollution is more critical than maintaining perfect accuracy at lower concentrations.

The key contrast between the two trained models lies in the distributional focus: the original dataset performs better on low-level $PM_{2.5}$ concentrations but struggles with extreme values. In comparison, the augmented dataset sacrifices some accuracy at lower concentrations to better capture the high-value events, which are crucial for understanding and managing pollution spikes. This trade-off is especially visible in the improvements in terms of RMSE, which penalizes large errors more severely. These results show that the model trained on augmented data is significantly better at predicting higher $PM_{2.5}$ values.

5. Discussion

The underestimation of high pollutant levels has been an issue frequently discussed in many studies [31]. This research addresses the challenge by applying data augmentation techniques before training the deep learning model. One of the key contributions of this study is the exploration of cluster-based undersampling, implemented at different cutoff thresholds and partial sampling ratios, which helped mitigate class imbalance and improve model performance. Our findings indicate that the higher cutoff threshold of 35.5 $\mu g/m^3$ resulted in superior model performance when compared to the lower threshold of 12.1 $\mu g/m^3$, as the 35.5 $\mu g/m^3$ threshold more effectively differentiated between low- and high-value points. The most optimal partial sampling ratio for the 35.5 $\mu g/m^3$ cutoff threshold was found to be 20/80. Previous studies, such as that of [49], explored data augmentation through linear interpolation to generate synthetic data and increase dataset size. Their approach significantly improved the performance of models like GRU and LSTM—yielding up to a 31% improvement in MAPE. However, the study primarily focused on increasing the overall volume of data without addressing class imbalance, which is a critical challenge in the prediction of extreme air pollution events. In contrast, [48] directly tackled dataset imbalance using random oversampling techniques to increase the representation of high-value samples. While their approach helped increase the representation of high-value samples, it led to overfitting on these samples and subsequently degraded the model's performance on the whole dataset. In contrast, our use of cluster-based undersampling allowed the model to avoid overfitting to high-value samples, resulting in improved prediction performance not only for the high-value samples but also for the dataset overall. These findings, however, align with [48] in highlighting the importance of partial sampling ratios, with their study identifying 30/70 as optimal for certain datasets and reinforcing the idea that fully balanced datasets are not always the best approach. Other studies, including that of [72], suggest that each dataset's unique characteristics necessitate tailored sampling strategies. In our case, the 20/80 ratio paired with the 35.5 $\mu g/m^3$ cutoff provided the best performance in capturing high-value points without over-suppressing the majority class, underscoring the importance of strategic undersampling for achieving balanced model generalization.

The results of this study not only highlight the effectiveness of cluster-based undersampling and tailored cutoff thresholds in improving $PM_{2.5}$ forecasting but also carry broader implications for air quality management. By addressing the frequent underestimation of high pollutant levels, our methodology contributes to more accurate identification of critical pollution episodes, which is essential for timely public health interventions. The superior performance achieved using a 35.5 $\mu g/m^3$ cutoff threshold underscores the importance of selecting thresholds that align with the dataset's characteristics and the targeted application. This finding suggests that air quality models must adopt region-specific or context-driven thresholds to ensure reliable predictions, especially when forecasting extreme pollution levels. Moreover, the partial sampling ratio of 20/80 demonstrates

that balancing the dataset does not necessarily mean achieving equal representation of classes; rather, an optimal balance must consider the distribution and nature of the data to maximize model performance.

Future work could enhance the model by incorporating additional data sources that influence $PM_{2.5}$ levels. Urban traffic data, which are crucial in accounting for emissions from vehicles, and industrial activity data from factories and power plants would provide more detailed insights into spikes in pollution. Including weather data such as wind patterns and forecasts could improve the model's accuracy in predicting pollutant dispersion across regions. In addition to data augmentation through cluster-based undersampling, more advanced techniques like Generative Adversarial Networks (GANs) could be explored to generate realistic synthetic data for extreme pollution events, which are rare but critical to forecast [77]. Another promising avenue would be extending the model to perform multistep predictions, forecasting $PM_{2.5}$ concentrations over multiple time steps rather than just the next step, which would be particularly valuable for air quality forecasting over longer periods like days or weeks. Moreover, extending the methodology to other pollutants, such as nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and ozone (O_3), would allow for a comprehensive air quality forecasting framework, enabling cities to predict and address multiple pollutants simultaneously. Given that many pollutants interact synergistically to exacerbate health effects, multi-pollutant models would enhance the precision of interventions. Additionally, applying this approach to different regions or urban areas would help validate the model's generalizability. Regional variations in pollution sources, meteorological factors, and population density may require adaptive strategies, such as incorporating localized data or adjusting the undersampling strategy to align with regional conditions.

From a methodological perspective, future extensions include the integration of multi-step forecasting, enabling predictions over longer time horizons. This would be particularly valuable for planning city-level interventions, such as scheduling traffic restrictions or industrial shutdowns during predicted high-pollution periods. The incorporation of additional data sources, such as traffic, industrial activity, and weather forecasts, could further enhance the model's robustness by capturing critical predictors of $PM_{2.5}$ variations. Lastly, advanced techniques like Generative Adversarial Networks (GANs) could be explored to synthesize data for rare but impactful extreme pollution events, addressing a key limitation of current datasets. These extensions would not only refine the methodology but also position it as a cornerstone for developing next-generation air quality forecasting systems.

6. Conclusions

This study demonstrates that the 35.5 $\mu g/m^3$ threshold consistently outperforms the 12.1 $\mu g/m^3$ threshold across key metrics like RMSE and R^2, likely due to its better representation of higher pollution values. The choice of partial sampling ratio proved crucial, with 50/50 optimal for the 12.1 $\mu g/m^3$ threshold and 20/80 optimal for the 35.5 $\mu g/m^3$ threshold, effectively balancing the need to capture both frequent and extreme pollution events. The model with the best performance (RMSE: 2.080, MAE: 1.386, R^2: 0.914) utilized the 35.5 $\mu g/m^3$ threshold and a 20/80 partial sampling ratio. Overall, models trained on resampled data significantly outperformed those trained on the original dataset, demonstrating the importance of data augmentation in handling imbalanced datasets and improving forecast accuracy, especially for high-value pollution scenarios.

The findings of this study have important practical implications. Accurate $PM_{2.5}$ forecasting is essential for timely public health interventions, particularly in urban areas prone to extreme pollution levels. By tailoring threshold selection and resampling strategies to the characteristics of the data, forecasting models can provide more reliable predictions,

enabling policymakers and city planners to take targeted actions to mitigate health and environmental risks.

Future research could build on these contributions by exploring additional thresholds tailored to specific regional air quality standards, ensuring broader applicability of the methodology. Incorporating supplementary data sources such as urban traffic patterns, industrial activity, and meteorological variables could further enhance the model's ability to capture the complex factors driving $PM_{2.5}$ fluctuations. Advanced techniques like Generative Adversarial Networks (GANs) could be employed to generate synthetic data for rare, extreme pollution events, addressing data scarcity challenges. Expanding the geographic scope of the model to include diverse regions and testing its performance with different pollutants such as NO_2, SO_2, and O_3 could create a comprehensive air quality forecasting system. Lastly, extending the model to perform multistep predictions would provide long-term forecasting capabilities, supporting more effective planning and intervention strategies over extended periods. These directions offer promising opportunities to refine and expand the impact of $PM_{2.5}$ forecasting models on air quality management for public health and disaster events such as wars and wildfires [78].

Author Contributions: Conceptualization, C.Y. and A.S.M.; methodology, P.P.; software, P.P.; validation, P.P.; formal analysis, P.P. and A.S.M.; investigation, P.P.; resources, C.Y.; data curation, A.S.M.; writing—original draft preparation, P.P. and A.S.M.; writing—review and editing, P.P., A.S.M., C.Y.; visualization, P.P. and A.S.M.; supervision, C.Y.; project administration, C.Y.; funding acquisition, C.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NASA AIST (NASA-AIST-QRS-23-02); NASA Goddard CISTO, and NSF I/UCRC program (1841520).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in this study are openly available from publicly accessible repositories. The AirNow PM2.5 data can be accessed via the AirNow API. The MODIS MAIAC AOD data are available from the NASA LP DAAC repository. The ERA5 reanalysis data are available from the Climate Data Store—Copernicus. The elevation data (GMTED2010) are available from the USGS GMTED2010 repository.

Conflicts of Interest: The authors declare no conflict of interest.

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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]

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

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. [CrossRef] [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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
47. Yen, S.J.; Lee, Y.S. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Syst. Appl.* **2009**, *36*, 5718–5727. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef] [PubMed]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
65. Lyapustin, A.; Wang, Y.; Korkin, S.; Huang, D. MODIS Collection 6 MAIAC algorithm. *Atmos. Meas. Tech.* **2018**, *11*, 5741–5765. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]
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. [CrossRef]