Air Quality Modelling with Satellite Data

Ayooluwa Agbato
aoa236@student.aru.ac.uk
Lorenzo Garbagna
lg673@pgr.aru.ac.uk
Lakshmi Babu Saheer
lakshmi.babu-saheer@aru.ac.uk
Mahdi Maktab Dar Oghaz
mahdi.maktabdar@aru.ac.uk

Faculty of Science and Engineering
Anglia Ruskin University
Cambridge, UK

Abstract

Air pollution poses significant environmental and public health challenges. Monitoring air quality is difficult, particularly in regions without ground-based stations. Satellite data provide a scalable alternative by offering consistent, publicly accessible environmental measurements. This research proposes a multimodal deep learning model for predicting tropospheric NO_2 concentrations using Sentinel-2 multispectral imagery, Sentinel-5P NO_2 data, and auxiliary atmospheric features such as population density, altitude, and land classification. The model employs a MobileNetV3 backbone for Sentinel-2 inputs, a compact CNN for Sentinel-5P, and a fully connected network for tabular features. These modalities are fused to predict pollutant concentrations. Trained on an extensive European dataset, the model achieved an R^2 score of 0.685 in its base form, improving to 0.73 with network customisation. Experiments show that combining all three data sources improves predictive accuracy, with Sentinel-2 contributing the most. The proposed model is computationally efficient, demonstrating potential for deployment in regions with limited processing resources.

1 Introduction

Air pollution is a major environmental and public health concern, contributing to respiratory illnesses, premature deaths, and climate change. The rise of industrialization and fossil fuel use has intensified pollution levels globally, emphasizing the need for accurate air quality monitoring. Traditional approaches rely on ground-based monitoring stations, which provide localized data but suffer from limited spatial coverage, high operational costs, and maintenance challenges. Satellite remote sensing offers a scalable alternative by providing atmospheric measurements at regional to global scales. Platforms such as Sentinel-2 and Sentinel-5P capture multispectral imagery and pollutant concentrations, enabling researchers to track emissions, analyze trends, and model air quality. However, satellite data are limited in spatial resolution (typically $\sim$1 km), making fine-grained pollutant estimation difficult. Machine learning (ML) techniques, particularly Convolutional Neural Networks (CNNs), have shown strong performance in image analysis and feature extraction. Deep CNNs, however, can be

computationally intensive, limiting their deployment in resource-constrained environments. Transfer learning, which leverages pre-trained models for specific tasks, has emerged as an effective approach to improve computational efficiency and predictive performance. Despite these advances, the application of multimodal CNNs to satellite-based air quality modeling remains underexplored. Integrating multi-sensor satellite data with ground-based measurements can potentially improve air quality predictions. However, existing models often fail to combine these heterogeneous data sources effectively or achieve high accuracy across diverse regions. There is a pressing need for scalable, accurate, and computationally efficient air quality models that can provide predictions in areas lacking monitoring infrastructure.

2 Related Works

Air quality has traditionally been measured using ground-based monitoring stations, with additional techniques introduced over time. A key challenge for these methods is the transport of air over large areas, which limits spatial coverage [4]. Remote sensing addresses this issue by enabling more accurate and spatially extensive measurements of pollutants. Satellites, in particular, provide rapid, global-scale observation capabilities and are often preferred over other platforms such as aircraft or drones due to their efficiency and data quality.

Early approaches to remote sensing-based air quality modeling include Land-Use Regression (LUR) and Kriging. LUR is a statistical method that estimates spatial variations in pollution by relating air quality measurements to land-use characteristics. A notable application was the Small Area Variations in Air quality and Health (SAVIAH) study, which linked traffic-related NO_2 concentrations over a two-week period using GIS data in the Netherlands [2]. Kriging, also known as Gaussian Process Regression, is a geostatistical interpolation technique that predicts values at unsampled locations based on surrounding measurements, providing both predictions and variance estimates [13].

However, these traditional approaches have limitations. LUR relies on GIS datasets, which vary in availability and granularity across regions, while Kriging depends on the density of ground stations. LUR models also struggle to separate the influence of multiple pollutants, often producing correlated predictions. Advances in machine learning (ML) using satellite data have addressed many of these issues. For instance, Young et al. [15] compared traditional methods with ML models using satellite inputs and observed improvements in predictive accuracy.

Initial ML applications for air quality using satellite imagery include Support Vector Machines (SVMs) to identify spatial patterns in remote sensing images. In this study, SVMs were compared with Maximum Likelihood Classifiers (MLC) and Decision Tree Classifiers (DTC). MLC was found to be limited by its assumption of normally distributed class signatures. The input data consisted of spatially degraded Thematic Mapper (TM) images with six spectral bands (1-5, 7) converted to top-of-atmosphere reflectance. Kernel selection played a critical role in SVM performance, with experiments showing that decision boundaries—and thus prediction accuracy—were highly sensitive to kernel type and parameters. SVMs ultimately outperformed both MLC and DTC approaches.

Deep learning (DL) has transformed air quality modeling, offering improved performance over traditional statistical methods [17]. Early approaches included Deep Belief Networks (DBNs), which capture hierarchical representations and perform well with limited labeled data [14]. However, DBNs are computationally expensive and often lack interpretability.

Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) architectures, model temporal dependencies effectively, improving prediction accuracy for short-term pollutant trends [3, 6, 7, 8, 12]. Hybrid frameworks combining LSTMs with CNNs, Autoencoders, or transfer learning further enhance performance and generalization across heterogeneous datasets.

Transformer models, using self-attention mechanisms, provide superior long-term temporal modeling and robustness to missing data. Models such as Informer and ResInformer demonstrate state-of-the-art performance for sequence-based pollutant prediction, though they require large datasets and high computational resources [1].

Convolutional Neural Networks (CNNs) excel at capturing spatial dependencies from satellite imagery [5, 11, 16, 18]. Early CNNs processed medium-resolution images, while hybrid and multimodal CNN architectures, combining Sentinel-2 and Sentinel-5P data with ground-based tabular features, have significantly improved pollutant prediction accuracy [9, 10]. CNNs, however, are limited in temporal modeling, motivating the use of CNN-RNN hybrids for spatiotemporal predictions.

Overall, DL-based models have evolved from DBNs to CNN, RNN/LSTM, and transformer architectures, each providing complementary strengths: DBNs for unsupervised feature learning, CNNs for spatial modeling, RNNs/LSTMs for temporal dynamics, and transformers for scalable sequence modeling. These advances have enabled more accurate, scalable, and multimodal air quality estimation across diverse regions.

3 Methodology

3.1 Dataset

The dataset utilized includes tabular data that exploits important information about the ground measurement centers such as altitude, population density station and area type as well as European Space Agency's (ESA) Copernicus missions Sentinel-2 and Sentinel-5P data during the 2018-2020 timeframe. Each satellite measurement, which is about 1.2 × 1.2km in size, covers about 3100 locations in Europe and about 100 on the US West Coast. The locations were meticulously chosen so that the measurement is centered at the location of an air quality measurement station on the ground, making it possible to analyze spatiotemporally aligned remote sensing and ground-based measurements The 12 Sentinel-2 bands have been cropped to 120 X 120 pixels and unsampled to a resolution of 10m. Some regions have multiple Sentine-2 images available. The images are grouped into folders based on their regions and are stored as binary NumPy ".npy" files. By linking measurements from successive satellite overpasses to a similar rectangular grid that spanned roughly 5X5km across Europe, the Sentinel-5P data was already pre-processed. The Sentinel-5P data was linearly interpolated to 10m resolution and cropped to 120X120 pixel around the locations of interest so that the Sentinel-5P data (5X3.5km, rescaled to 5X5km) could be in the same type as the Sentinel-2 (10m to 60m, upscaled to 10m) imaging resolutions. In accordance to ESA recommendations, all measurements having a QA flag (qa value) below 75 were eliminated. The Sentinel-5P data are arranged based on locations and stored as '.netcdf' files. Three of these files, containing Sentinel-5P measurements at various temporal frequencies (2018-2020, quarterly and monthly) are provided for each location. Eight text/numerical columns made up the tabular data which were altitude (numerical), population density (numerical), binary encoded features for rural, suburban or urban area and binary encoded features for

whether the station was traffic, industrial or background monitoring station.

3.2 Preprocessing

The multimodal dataset was prepared through a structured preprocessing pipeline. First, tabular descriptors and pollutant values were read from a central CSV and linked with corresponding Sentinel-2 and Sentinel-5P files, with missing NO_2 entries discarded. To reduce I/O overhead, Sentinel-5P .netcdf files were indexed by station ID, and each sample was stored as a multimodal dictionary containing tabular attributes, Sentinel-5P maps, Sentinel-2 imagery, and metadata. Sentinel-2 images were reformatted from (H,W,C) to (C,H,W), cropped to 120×120 pixels, and reordered to match BigEarthNet band conventions. All modalities, including Sentinel-2 bands, Sentinel-5P inputs, tabular features, and pollutant targets, were normalized using precomputed global means and standard deviations, with inverse transforms applied for interpretability of predictions. Data augmentation (random flips and 90° rotations) was applied to Sentinel-2 and Sentinel-5P inputs to improve robustness. Finally, NumPy arrays were converted to PyTorch tensors to enable GPU acceleration. The dataset of 1,318 monitoring stations was split into 70% training, 15% validation, and 15% testing, stratified by station ID to avoid location bias.

3.3 Model architecture

We propose several enhancements to the base AQNet architecture that integrates multispectral Sentinel-2 imagery, tropospheric NO_2 maps from Sentinel-5P, and station-level tabular data. The design builds on established CNN backbones but introduces domain-specific adaptations to improve efficiency and prediction accuracy.

Sentinel-2 Backbone. Sentinel-2 inputs consist of 12 spectral bands cropped to 120×120 pixels and harmonized to 10 m resolution. A MobileNetV3 backbone was selected for its efficiency in lightweight vision tasks. The input layer was modified to accept 12 channels instead of RGB, and the classification head was replaced with a 640-dimensional feature extractor suitable for regression.

Sentinel-5P Backbone. Tropospheric NO_2 inputs are processed by a compact CNN designed for low-resolution, single-channel data. The network consists of two convolutional layers (1→10 and 10→15 channels) with ReLU activations and max pooling, followed by a fully connected layer producing a 128-dimensional feature vector. This lightweight design balances speed and expressiveness.

Tabular Backbone. Eight structured features (altitude, population density, and binary encodings of area type and station type) are processed by a three-layer feedforward network (8→16→32→32 dimensions), yielding a compact 32-dimensional representation of contextual attributes.

Fusion and Regression. Sentinel-2 and Sentinel-5P features are concatenated and passed through a three-layer "satellite head" (768→384→192→96 dimensions) with ReLU activations, compressing the joint representation. This output is fused with the 32-dimensional tabular vector to form a 128-dimensional feature, which is passed through a regression head with dropout regularization to produce the final pollutant prediction.

Fine-tuning. Several architectural variations were explored. The inclusion of a Squeeze-and-Excitation (SE) block in the Sentinel-5P backbone improved performance by reweighting channel importance. Replacing ReLU with GELU activations further enhanced accuracy at the cost of higher computation. The final configuration provided a balance of predictive

power and efficiency, suitable for large-scale deployment. SE blocks recalibrate feature maps by learning channel-wise attention weights, so that the model can give more importance to pollutant-relevant spectral and spatial features, while suppressing less informative signals. This mechanism enhances multimodal fusion by aligning the importance of different feature channels. On the other hand, for AQNet's multimodal architecture and fusion head, using GELU provides smooth, probabilistic gating that preserves informative small and negative signals across modalities, improves gradient flow, lowers training variance, and often outperforms ReLU in modern LayerNorm-based setups, providing more stable fine-tuning.

4 Results

4.1 Base Model

The original AQNet model was developed as a multimodal deep learning framework integrating Sentinel-2 imagery, Sentinel-5P atmospheric data, and tabular ground-level data. The base model employed a MobileNetV3-Small backbone, a compact CNN for Sentinel-5P, and a three-layer MLP for auxiliary features. These streams were fused in a regression head to predict NO_2 concentrations. Table 1 shows the results of the baseline AQNet model.

Across five randomized runs, the model demonstrated consistent performance. The mean R^2 score was 0.685 (range: 0.662–0.703), with MAE between 3.53–4.02 $\mu g/m^3$ and MSE between 22.07–33.52. Despite its lightweight design, the model exhibited robust convergence behaviour, providing a strong baseline. Compared with Random Forest and XGBoost baselines for tabular, AQNet achieves lower MAE (average 3.70 against 8.1 for XGBoost and 8.7 for Random Forest) and higher R^2 (average 0.68 against 0.62 and 0.58).

Table 1: Performance of the base AQNet model across five runs.

Run	R^2	MAE	MSE
1	0.695	3.621	22.074
2	0.666	4.021	33.512
3	0.702	3.534	22.774
4	0.699	3.762	24.748
5	0.662	3.601	25.182

4.2 Impact of Model Customisation

Several enhancements were introduced, including SE attention blocks, GELU activations, and refined Sentinel-2/5P backbones.

Base vs. Fine-tuned Model Performance Table 2 presents the results of the described model. The fine-tuned AQNet outperformed the baseline, achieving a mean R^2 of 0.728 (best run: 0.755), with reduced MAE (3.31 $\mu g/m^3$) and lower MSE (down to 18.4). The training and validation curves showed stable improvement without significant overfitting.

Computational Efficiency Despite added complexity, runtime efficiency was preserved. MobileNetV3 and SE blocks enabled lightweight computation, with only marginal increases in training time. GPU memory usage remained within limits, and early stopping ensured most runs terminated well before the 30-epoch cap.

Table 2: Performance of the fine-tuned AQNet model.

Run	R^2	MAE	MSE
1	0.739	3.31	19.2
2	0.744	3.21	18.4
3	0.740	3.35	18.6
4	0.715	3.43	18.9
5	0.720	3.32	18.7

4.3 Ablation Study

In order to evaluate which component of the fine-tuned AQNet contributed the most to the performances of the model, an ablation study has been performed. The architecture has been systematically tested by removing components to validate the importance on each on the total accuracy of the model. Table 3 shows the results of the ablation.

Table 3: Ablation of fine-tuned AQNet.

Model Variant	R^2
Without Sentinel-2	0.63
Without Sentinel-5P	0.70
Without Tabular Data	0.68
Single Modality - Sentinel-5P	0.59
Single Modality - Tabular only	0.61
Fine-tuned AQNet	***0.73***

The fine-tuned AQNet achieves the best performance (R^2 = 0.73). Removing Sentinel-2 produces the largest drop (from 0.73 to 0.63), while removing Sentinel-5P results in a smaller decline (from 0.73 to 0.70). Excluding the tabular branch lowers performance to 0.68. These trends suggest that Sentinel-2 contributes the most to the model's accuracy, likely due to its higher spatial resolution and richer texture and spectral cues. Tabular features provide important contextual priors that help stabilise predictions, especially when remote-sensing inputs are ambiguous, while Sentinel-5P adds complementary information: its benefit is modest in isolation but becomes more useful when combined with S2.

Single-modality baselines reinforce this interpretation: tabular alone reaches an R^2 of 0.61, outperforming S5P (0.59). Furthermore, S2+S5P improves over tabular (0.68 against 0.61), while S5P+Tabular only reaches the tabular-only baseline (0.61), implying that S5P coarse signal is not sufficiently leveraged without the fine-grained structure from S2. The full model improves from the S2+S5P variant (0.68) to 0.73, highlighting that the fusion extracts patterns that each branch cannot provide alone.

S2 high-resolution spatial patterns likely drive the strongest gains due to how they encode fine-grained texture, edges, and morphological cues that can discern between land-cover types and local structures. These details capture neighbourhood context and boundary information that coarse signals can miss, enabling the model to learn sharper associations between surface features and target variables. Rich spatial variability also reduces ambiguity in homogeneous regions, improving discrimination and calibration. In multimodal fusion, such spatial structure provides an anchor for aligning and interpreting complementary inputs, allowing the network to exploit their relationships with greater precision.

5 Future Work and Application

These experiments highlight the effectiveness of multimodal architectures for air quality modelling. Beginning with a strong baseline, architectural refinements boosted predictive accuracy without major computational costs. for real world applications, AQNet's multimodal design allows scalable deployment in cities where monitoring stations are sparse. For instance, urban planners could use this fine-tuned AQNet architecture outputs to identify traffic corridors, informing green infrastructure placement. Furthermore, public health agencies could utilise the model for early-warning systems, delivering targeted alerts to vulnerable populations. The lightweight MobileNet backbone further enables deployment on edge devices, making the architecture suitable for real-time applications in developing regions.

Future work could extend AQNet by feeding fused representations into an RNN or Transformer. This would enable modeling lag effects, seasonal variations, and pollutant transport dynamics. Furthemore, extending to multi-output regression for multiple pollutants, and adapting the model for real-time, large-scale forecasting are also additional paths to improve on the architecuture and expand the scope of the model for real world applications and air quality monitoring systems.

6 Conclusion

This work evaluated and extended AQNet, a multimodal deep learning framework for predicting NO_2 concentrations from Sentinel-2 imagery, Sentinel-5P atmospheric data, and auxiliary ground-level features. Using MobileNetV3 backbones and lightweight convolutional layers, the study demonstrated the effectiveness of fusing heterogeneous data streams for scalable air quality modelling.

The baseline AQNet achieved a mean R^2 of 0.685 with low error margins, validating its ability to capture spatial and atmospheric signals. Architectural enhancements—including SE attention blocks, GELU activations, and improved backbone configurations—pushed performance beyond $R^2 = 0.73$, with accuracy gains most pronounced in the Sentinel-2 processing stream. Importantly, these improvements were achieved with minimal increases in computational cost, preserving efficiency while enhancing predictive strength. These findings align with sustainable urban development goals, as fine-tuned AQNet provides a scalable solution for air quality monitoring in regions lacking dense sensor networks. By integrating remote sensing and ground data, the model supports data-driven policies in urban planning, environmental justice, and public health resilience.

Overall, the project demonstrated the potential of multimodal neural architectures for accurate and lightweight air quality prediction, providing a foundation for future extensions such as temporal modelling, transformer-based attention, and multi-pollutant forecasting.

References

[1] M. A. A. Al-qaness, A. Dahou, A. A. Ewees, L. Abualigah, J. Huai, M. Abd Elaziz, and A. M. Helmi. Resinformer: Residual transformer-based artificial time-series forecasting model for pm2.5 concentration in three major chinese cities. *Mathematics*, 11 (2):476, 2023. doi: 10.3390/math11020476.

[2] D. J. Briggs, S. Collins, P. Elliott, P. Fischer, S. Kingham, E. Lebret, K. Pryl, H. Van Reeuwijk, K. Smallbone, and A. Van Der Veen. Mapping urban air pollution using gis: A regression-based approach. *International Journal of Geographical Information Science*, 11(7):699–718, 1997. doi: 10.1080/136588197242158.

[3] W. Ding and Y. Zhu. Prediction of pm2.5 concentration in ningxia hui autonomous region based on pca-attention-lstm. *Atmosphere*, 13(9):1444, 2022. doi: 10.3390/atmos13091444.

[4] O. Hertel, J. Salmond, W. Bloss, I. Salma, S. Vardoulakis, R. Maynard, M. Williams, and M. E. Goodsite. *Air Quality in Urban Environments*. The Royal Society of Chemistry, 2009. doi: 10.1039/9781847559654.

[5] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi. Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research*, 23(22):22408–22417, 2016. doi: 10.1007/s11356-016-7812-9.

[6] X. Li, L. Peng, X. Yao, S. Cui, Y. Hu, C. You, and T. Chi. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental Pollution*, 231:997–1004, 2017. doi: 10.1016/j.envpol.2017.08.114.

[7] J. Ma, Y. Ding, V. J. L. Gan, C. Lin, and Z. Wan. Spatiotemporal prediction of pm2.5 concentrations at different time granularities using idw-blstm. *IEEE Access*, 7:107897–107907, 2019. doi: 10.1109/ACCESS.2019.2932445.

[8] A. G. Mengara Mengara, E. Park, J. Jang, and Y. Yoo. Attention-based distributed deep learning model for air quality forecasting. *Sustainability*, 14(6):3269, 2022. doi: 10.3390/su14063269.

[9] A. Rowley and O. Karakuş. Predicting air quality via multimodal ai and satellite imagery. *Remote Sensing of Environment*, 293, 2023. doi: 10.1016/j.rse.2023.113609.

[10] L. Scheibenreif, M. Mommert, and D. Borth. Estimation of air pollution with remote sensing data: Revealing greenhouse gas emissions from space, 2021. URL http://arxiv.org/abs/2108.13902.

[11] A. Sharma, X. Liu, X. Yang, and D. Shi. A patch-based convolutional neural network for remote sensing image classification. *Neural Networks*, 95:19–28, 2017. doi: 10.1016/j.neunet.2017.07.017.

[12] X. Sun and W. Xu. Deep random subspace learning: A spatial-temporal modeling approach for air quality prediction. *Atmosphere*, 10(9):560, 2019. doi: 10.3390/atmos10090560.

[13] V. van Zoest, F. B. Osei, G. Hoek, and A. Stein. Spatio-temporal regression kriging for modelling urban no2 concentrations. *International Journal of Geographical Information Science*, 34(5):851–865, 2020. doi: 10.1080/13658816.2019.1667501.

[14] X. Yan, Z. Zang, N. Luo, Y. Jiang, and Z. Li. New interpretable deep learning model to monitor real-time pm2.5 concentrations from satellite data. *Environment International*, 144, 2020. doi: 10.1016/j.envint.2020.106060.

[15] M. T. Young, M. J. Bechle, P. D. Sampson, A. A. Szpiro, J. D. Marshall, L. Sheppard, and J. D. Kaufman. Satellite-based no2 and model validation in a national prediction model based on universal kriging and land-use regression. *Environmental Science Technology*, 50(7):3686–3694, 2016. doi: 10.1021/acs.est.5b05099.

[16] K. Zhang, X. Yang, H. Cao, J. Thé, Z. Tan, and H. Yu. Multi-step forecast of pm2.5 and pm10 concentrations using convolutional neural network integrated with spatial–temporal attention and residual learning. *Environment International*, 171:107691, 2023. doi: 10.1016/j.envint.2022.107691.

[17] S. Zhou, W. Wang, L. Zhu, Q. Qiao, and Y. Kang. Deep-learning architecture for pm2.5 concentration prediction: A review. *Environmental Science and Ecotechnology*, 21, 2024. doi: 10.1016/j.ese.2024.100400.

[18] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017. doi: 10.1109/MGRS.2017.2762307.