Foundation Models for Mapping Emission Sources and Acute Respiratory Infection (ARI) Hotspots

Usman Nazir **Sara Khalid**
usman.nazir@ndorms.ox.ac.uk sara.khalid@ndorms.ox.ac.uk

Abstract

This study presents a cross-disciplinary approach combining foundation models, satellite imagery, and demographic health surveys to analyze the correlation between unregulated industrial activity and Acute Respiratory Infections (ARI) in South Asia. We detect brick kiln and factory chimneys using a sequential transformer chaining method and a multispectral Chimney Index. These detections are spatially joined with a geocoded dataset of ARI incidence developed from DHS surveys across India, Pakistan, and Bangladesh. Our findings reveal a statistically significant correlation between chimney density and ARI cases in children under five, underscoring the urgent need for regulatory and health interventions in high-emission zones.

1 Pathway to Climate Impact

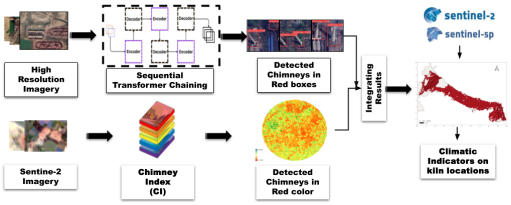
Acute Respiratory Infections (ARIs) remain one of the leading causes of morbidity and mortality in low- and middle-income countries (LMICs), especially among children. Previous studies have linked the high prevalence of ARI with exposure to air pollutants such as sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and carbon monoxide (CO), frequently emitted by unregulated brick kilns and industrial chimneys [1, 2, 3, 4]. These industrial sources not only release health-damaging pollutants but also contribute to climate change through emissions of black carbon and greenhouse gases, exacerbating local warming and air quality degradation. Thus, chimney smoke embodies a dual burden—driving both respiratory disease and regional climate stress. In this work, we present a pipeline that uses foundation models and satellite imagery to detect pollution sources and correlate them with DHS health data.

2 Introduction

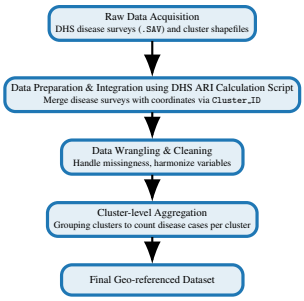
The Global Burden of Disease (GBD) framework has brought renewed attention to the worldwide impact of infectious and non-communicable diseases. Although GBD provides a macrolevel view, crucial gaps persist in the surveillance of Acute Respiratory Infections (ARI), particularly in low- and middle-income countries [5]. ARIs remain a leading cause of morbidity and mortality in these regions [6]. Many studies have shown that environmental factors, such as air pollution from factory chimneys and climate variability, critically affect the prevalence and outcomes of diseases [1].

Despite the recognized burden, robust, high-resolution data linking ARI incidence to specific environmental parameters remain scarce. Large-scale data collection efforts like the Demographic and Health Surveys (DHS) have generated valuable insights; however, these surveys are often underutilized for in-depth spatial epidemiology due to challenges in data extraction, geocoding, and integration with external environmental layers [7]. The Integrated Burden of Disease and Environmental Monitoring System (IBDEM) aims to remedy these challenges by offering a framework that merges DHS data, complete with cluster-level latitude/longitude coordinates—with environmental and demographic indicators.

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.



(a) Detection of emission sources pipeline.



(b) ARI geocoded data extraction pipeline.

Figure 1: End-to-end pipeline for mapping industrial emissions and respiratory health risk. (a) Chimney detection using spectral indices and transformers. (b) ARI data extraction and geo-merging pipeline.

Previous studies have also utilized data from MODIS, and Sentinel-5P to monitor air quality and its effects on respiratory health [3, 4]. Others have integrated DHS data for spatial epidemiology [2]. We build upon our earlier work [7] where we used transformer models for chimney detection, now incorporating health outcome datasets.

3 Data Sources

Chimney Detection Data: High-resolution satellite imagery for South Asia was obtained using Google Earth Downloader[1]. From this imagery, a multispectral Chimney Index (CI) was derived using Sentinel-2 bands to enhance the detection of industrial sites and brick kilns.

ARI Data: Acute Respiratory Infection (ARI) indicators and geospatial cluster coordinates were extracted from the Demographic and Health Surveys (DHS)[2]. Data cleaning and geo-merging were conducted following the DHS ARI Calculation script [3], ensuring harmonization of survey variations.

4 Chimney Detection Pipeline

A novel pipeline was developed for detecting brick kilns and industrial chimneys by combining low- and high-resolution imagery with spectral indices and transformer-based models. Schematic overview of these steps is shown in Figure 1 (a).

4.1 Chimneys detection using novel multi-spectral Chimney Index (CI)

A novel Chimney Index (CI) was developed using Sentinel-2 multi-spectral imagery to enhance the detection of industrial sites. It combines the Normalized Difference Vegetation Index (NDVI), Burn Index (BI), and Built-Up Index (BUI) with weighted contributions:

$$CI = (\omega_1 \times (1 - NDVI)) + (\omega_2 \times BI) + (\omega_3 \times BUI) \quad (1)$$

Low NDVI values highlight non-vegetated regions, high BI indicates heat-affected areas, and high BUI identifies urban zones. Their integration enables robust large-scale detection of chimneys and industrial structures in South Asia, where emissions significantly impact respiratory health.

4.2 Chimneys detection using sequential transformer chaining mechanism

A sequential filtering pipeline was applied to high-resolution satellite imagery of urban areas. First, a Vision Transformer performed broad classification with multiple prompts and a low confidence threshold. Next, Remote CLIP was applied with higher thresholds to refine chimney detection, followed by an additional filtering step targeting smokestacks. Manual review finalized the dataset.

[1] https://www.allmapsoft.com/geid/

[2] https://dhsprogram.com/

[3] https://github.com/DHSProgram/DHS-Indicators-R/blob/main/Chap10_CH/CH_DIAR.R

This multi-stage chaining of foundation models with human validation produced a precise and reliable inventory of industrial structures.

5 Acute Respiratory Infection (ARI) Geocoded Data Extraction Pipeline

DHS survey data were decoded, harmonized, and geo-merged with cluster coordinates to construct ARI incidence datasets at multiple spatial levels. Schematic overview is illustrated in Figure 1 (b).

5.1 Data Pre-processing and Integration

DHS datasets in `.SAV` format were decoded using official scripts to retain variable labels [8]. ARI-related indicators (e.g., cough, difficulty breathing, confirmed cases) were extracted and converted to CSV for analysis in Python and R. Geospatial files (shapefiles and cluster coordinates) were standardized to the WGS84 system and merged with survey data using `Cluster_ID` keys. Variables were harmonized across countries (e.g., `CH_ARI`, `NO_ARI_CASES`, `LATNUM`, `LONGNUM`), and records with extensive missing or inconsistent values were removed, while limited gaps ($<5\%$) were imputed.

5.2 Quality Control and Validation

Aggregate ARI incidence was cross-checked against DHS and WHO statistics [6], and random spot checks ensured geospatial accuracy and minimized duplication.

5.3 Dataset Composition

The final dataset spans over 30 LMICs (primarily Sub-Saharan Africa and South & Southeast Asia), with several thousand geocoded clusters. Each entry includes ARI incidence, geospatial coordinates, administrative boundaries, and survey year. Approximately 10% of clusters reported zero ARI cases, validated as either low-risk or reflecting small sample sizes.

Table 1: Correlation between ARI cases and environmental indicators. Top-2 correlations are highlighted.

Variable	CO_mean	NO_2_mean	SO_2_mean	Mean Temp. (°C)
NO_ARI_CASES	0.005	**0.031**	**0.170**	0.001

Table 2: Performance of chimney detection methods. Top-2 accuracies are highlighted.

Method / Stage	Accuracy (%)
Chimney Index (threshold = 0.4)	
Patparganj Industrial Area, Delhi (India)	79.17
Sundar Industrial Estate, Lahore (Pakistan)	68.42
Tongi Industrial Estate, Dhaka (Bangladesh)	70.00
Sequential Transformer Chaining	
Initial Vision Transformer Filtering	60.00
Secondary Remote CLIP Filtering	**85.00**
Tertiary Remote CLIP Refinement	**95.00**

Table 3: ARI cases prediction using multiple climatic and air-quality indicators.

Model	Loss Function	Validation MSE	Notes
BERT (baseline)	MSE	0.860	smaller architecture
TabTransformer (Our implementation)	MSE	0.145	larger model (dim=256, depth=6, heads=32)

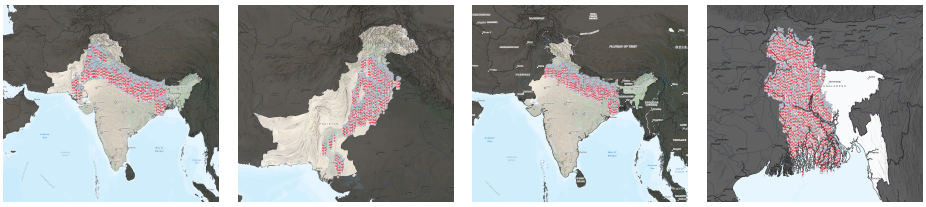
6 Results

We first examined the Pearson's correlations (see Table 1) between Acute Respiratory Infection (ARI) incidence and environmental indicators (CO, NO_2, SO_2, and mean temperature). Overall associations were weak, with SO_2 showing the strongest positive correlation ($r = 0.17$), while CO ($r = 0.005$), NO_2 ($r = 0.031$), and mean temperature ($r = 0.001$) showed negligible relationships.

Brick kilns and industrial chimneys in South Asia frequently burn coal, biomass, or low-grade furnace oil, which release SO_2 when sulfur is oxidized during combustion. As a major air pollutant, SO_2 irritates the respiratory tract, contributes to childhood ARI, and forms secondary pollutants such as sulfate aerosols and acid rain. These findings suggest pollutant-specific effects, with sulfur dioxide emerging as a more relevant driver of localized ARI burden.

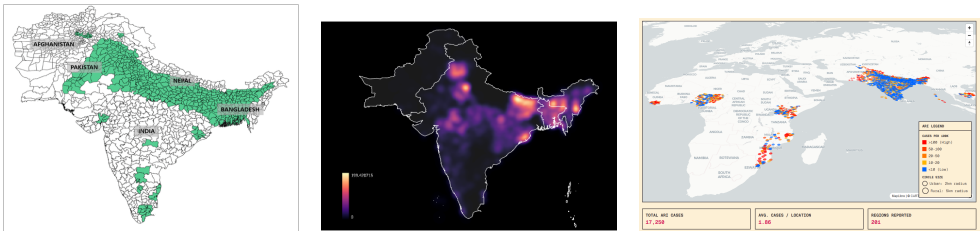
Figure 2 presents the geospatial distribution of brick kiln emissions (∼55,387 chimneys), and Figure 3 shows ARI incidence across South Asia, highlighting overlapping spatial hotspots of industrial activity and respiratory health burden. Quantitative evaluation in Tables 2, further demonstrates that the Chimney Index achieved regional accuracies of 68–79%, refined to 95% through Sequential Transformer Chaining [7].

For ARI prediction, we compared two transformer-based approaches. The TabTransformer [9] achieved a substantially lower validation error (MSE = 0.145) than the baseline BERT model [10], confirming its effectiveness for structured environmental data (see Table 3). The model took climatic and environmental indicators as inputs and predicted ARI case counts through a regression head built on top of embedding layers, multi-head self-attention, and feed-forward blocks.



(a) (b) (c) (d)

Figure 2: Geospatial visualization of emission sources across South Asia. (a) Brick Kiln Belt of South Asia (55,387 chimneys), (b) Kiln locations in Pakistan, (c) Kiln locations in India, and (d) Kiln locations in Bangladesh.



(a) (b) (c)

Figure 3: Geospatial visualization Acute Respiratory Infection (ARI) incidence across South Asia. (a) Brick Kiln Belt of South Asia (55,387 chimneys) (b) Raw DHS-reported ARI cases at cluster level. (c) Normalized ARI incidence per 100,000 population showing spatial hotspots over 30 low- and middle-income nations, mainly in Sub-Saharan Africa and South & Southeast Asia

Our findings validate the hypothesis that industrial air pollution sources are significantly correlated with respiratory health burdens in children. While the DHS spatial jitter limits absolute precision, the correlation patterns hold across three countries of South Asia. Further research should include time-series modeling and pollutant-level validation with Sentinel-5P.

7 Conclusion

This work demonstrates a scalable, data-driven framework for quantifying the health impacts of industrial emissions by integrating remote sensing, foundation models, and geocoded health surveys. By linking chimney detection with respiratory health outcomes, we provide new evidence on pollutant-specific risks, particularly the role of sulfur dioxide in driving Acute Respiratory Infection (ARI) burden. The approach not only advances spatial epidemiology but also supports the design of targeted regulatory and health interventions in high-risk urban areas. Looking forward, the framework can be extended with time-series analysis and additional environmental indicators to strengthen predictive modeling and inform climate and health policy.

References

[1] Godson Kalipe, Vikas Gautham, and Rajat Kumar Behera. Predicting malarial outbreak using machine learning and deep learning approach: a review and analysis. In *2018 international conference on information technology (ICIT)*, pages 33–38. IEEE, 2018.

[2] Zinabu Bekele Tadese, Debela Tsegaye Hailu, Aschale Wubete Abebe, Shimels Derso Kebede, Agmasie Damtew Walle, Beminate Lemma Seifu, and Teshome Demis Nimani. Interpretable prediction of acute respiratory infection disease among under-five children in ethiopia using ensemble machine learning and shapley additive explanations (shap). *Digital Health*, 10:20552076241272739, 2024.

[3] Sathya Swarup Aithal, Ishaan Sachdeva, and Om P Kurmi. Air quality and respiratory health in children. *Breathe*, 19(2), 2023.

[4] Shabana Siddique, Manas R Ray, and Twisha Lahiri. Effects of air pollution on the respiratory health of children: a study in the capital city of india. *Air Quality, Atmosphere & Health*, 4(2):95–102, 2011.

[5] Martin Sudmanns, Dirk Tiede, Stefan Lang, Helena Bergstedt, Georg Trost, Hannah Augustin, Andrea Baraldi, and Thomas Blaschke. Big earth data: disruptive changes in earth observation data management and analysis? *International Journal of Digital Earth*, 13(7):832–850, 2020.

[6] World Health Organization. Environmental risk factors and the burden of disease. `https://www.who.int/about/accountability/results/2018-2019`, 2018. Accessed: 2025-08-05.

[7] Hafiz Muhammad Abubakar, Raahim Arbaz, Hasnain Ahmad, Mubasher Nazir, and Usman Nazir. Mapping air pollution sources with sequential transformer chaining: A case study in south asia. In *NeurIPS Workshop on Tackling Climate Change with Machine Learning*, 2024. NeurIPS CCAI Workshop.

[8] Demographic and Health Surveys (DHS). Demographic and health surveys program. `http://www.dhsprogram.com`, 2025. Accessed: 2025-08-05.

[9] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.