Imperial College London
Department of Earth Science and Engineering
MSc Environmental Data Science and Machine Learning

Independent Research Project
Project Plan

Predicting Individual Physiological Responses to Pollution Using Transformer–Based Time-Series Models

by

Davide Baino

Email: davide.baino24@imperial.ac.uk

GitHub repository: https://github.com/ese-ada-lovelace-2024/irp-db24

Supervisors:
Dr. Christopher Pain
Dr. Boyang Chen

June 14, 2025

Table of Contents

1. Abstract

Air pollution remains a major global health and environmental concern, contributing to an estimated seven million deaths annually through the combined effects of outdoor and household exposure (WHO, 2025)[1]. While pollution levels are projected to decline, the ongoing impacts of climate change continue to pose serious risks. Simultaneously, advancements in wearable sensor technologies allow for the systematic collection of high-resolution physiological data over long periods of time (Roos & Slavich, 2023)[2].

This study aims to develop an identity map linking varying levels of air pollution to individual physiological responses. Such a framework will enable the prediction of health responses to pollution exposure, facilitating early warnings and personalised health recommendations. To achieve this, we propose a two-model approach: an initial general model to capture population-wide temporal trends, and a personalised one fine-tuned to individual characteristics. Together, these models will enhance the precision of forecasting and contribute to more effective, data-driven health interventions when reacting to a polluted environment.

2. Problem Description

2.1 Rationale and Literature Review

Air pollution represents a critical challenge in the 21st century, with significant implications for human health. For example, He et al. [3] estimate that air pollution reduces average life expectancy by 1.8 years worldwide and up to 3 years in highly polluted regions of China.

Pollution occurs when substances from human, biological, or natural sources enter the atmosphere at concentrations beyond typical levels, posing short- or long-term risks (Bernasconi, Angelucci, & Aliverti, 2022)[4]. Pollutants are categorized as either primary (directly emitted, such as PM, CO, and NO) or secondary (formed through chemical reactions, like O and NO, often found far

from their original sources). This study primarily examines criteria pollutants, particularly PM10 and PM2.5, due to their severe health risks (Bernasconi, Angelucci, & Aliverti, 2022)[4].

Historically, air quality has been monitored using fixed-location stations, providing aggregated data at a city or regional level. While useful for broad trends, this overlooks personal exposure, which varies with location, activity, and individual health (Hu et al., 2014)[5]. Population averages can mask true personal risk.

Recent research has improved pollution forecasting, yet gaps remain in linking these predictions to health outcomes. For example, the Breath study employs a transformer-based model to predict NO levels in India with high accuracy (Verma et al., 2024)[6]. However, it does not explore how these pollution fluctuations affect individual or population health, limiting its utility for policymaking or preventative healthcare.

In contrast, Atzeni et al. (2025)[7] developed a machine learning pipeline for short-term respiratory disease prediction. Their work underscores the importance of stratifying individuals before modelling and demonstrates the effectiveness of traditional methods like Logistic Regression, Random Forest, and XGBoost. Nevertheless, it does not leverage modern deep learning techniques for time-series analysis, which could better capture temporal patterns in physiological data.

2.2 Objectives

The approach proposed in this study directly addresses the challenge outlined in the BEHRT initiative by integrating IoT-enabled wearable devices to support proactive and personalised health interventions Li, Y. et al. (2020)[8]. To this end, a transformer-based time-series deep learning architecture is proposed. The first objective is to harness the transformer's strength in modelling long-range temporal dependencies to identify population-level trends, thereby enabling the construction of an identity map that links pollution exposure to physiological responses. The second objective is to deliver real-time, individualised insights—alerting users to expect physiological changes when encountering similar pollution levels in the future.

3. Dataset

The datasets used in this study come from two distinct sources, offering complementary information to investigate the relationship between air pollution and individual health responses.

The first dataset is provided by the INHALE project and consists of data from 59 participants aged between 20 and 75 years, including 33 non-asthmatic and 26 asthmatic individuals. Each participant was equipped with wearable sensors that recorded information on air pollution exposure, respiratory health, and physical activity. Although the temporal coverage of the INHALE dataset is

limited, two-week summer and winter periods, it provides a more targeted view of personalised exposure and associated health outcomes over different time periods, as shown in figure 1.

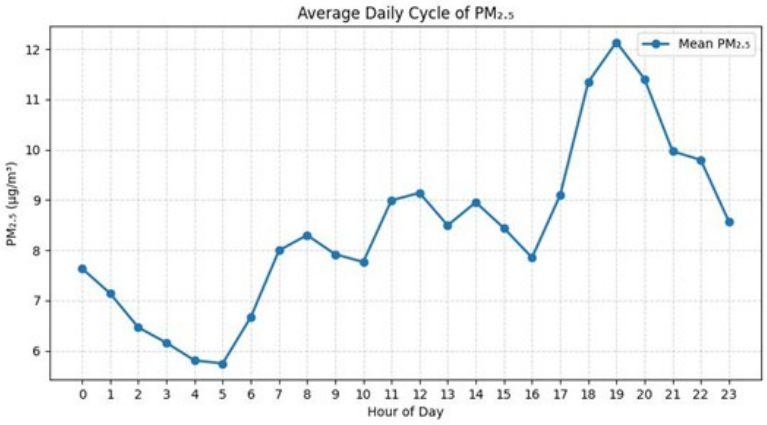

Figure 1: Hourly PM. levels from the INHALE dataset, showing clear peaks in individual exposure between 18:00 and 21:00.

The second dataset comes from the OpenWeather API and provides real-time air pollution levels, including key pollutants such as PM2.5 and NO2. Crucially, this dataset is geolocated using GPS coordinates, offering pollution data at a spatial resolution of approximately 200 metres. When matched with the location data collected through INHALE, this enables a spatially-aware analysis of personal exposure to air pollution, allowing for the integration of environmental data with individual physiological responses. Although a 200-metre resolution may not capture highly localised variations in pollution, this limitation is mitigated by the more granular exposure data available directly from the INHALE dataset.

4. Methodology

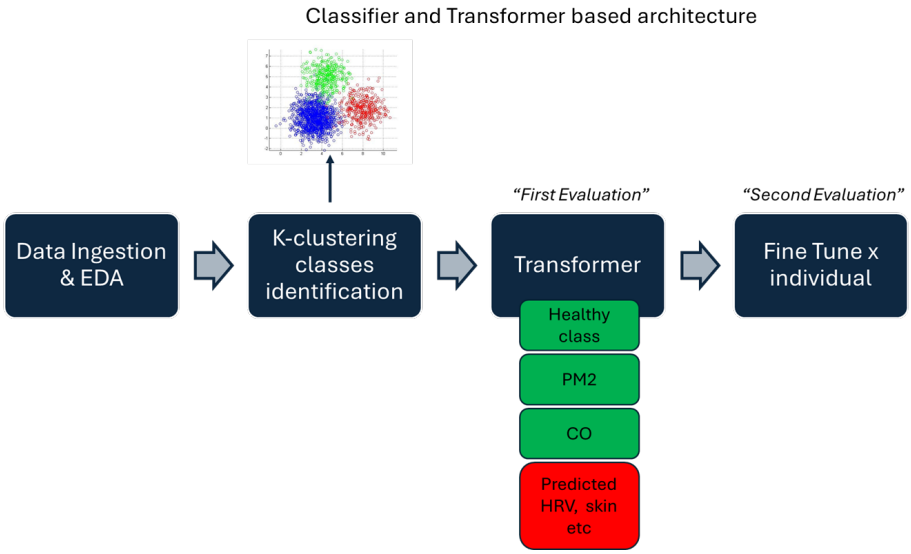

Figure 2: General transformer-based architecture, trained on OpenWeather and INHALE architecture and subsequently fine-tuned at individual level.

The analysis begins with ingesting and exploring both datasets through comprehensive exploratory data analysis (EDA) to examine feature distributions, scales, and relationships. Given the heterogeneity of variables—ranging from physiological metrics to pollution indicators—normalisation will standardise scales for comparability. Missing values will be imputed using time-series techniques (e.g., ARIMA for temporal gaps) and Multiple Imputation by Chained Equations (MICE) to address statistical missingness (Royston & White, 2011)[12]. Outliers, particularly those arising from wearable sensor noise, will be identified and corrected to ensure high level of data quality.

To interpret feature relevance, SHAP (Shapley Additive Explanations) values will be computed to quantify the contribution of each variable to model predictions (Lundberg & Lee, 2017)[9]. Following data cleaning and transformation, an initial unsupervised clustering approach (e.g., k-means) will group individuals based on their responses to exposure to pollution(AlNuaimi & AlBaldawi, 2024)[11] . This acknowledge the heterogeneity of how individuals react to similar pollution level.

The resulting cluster labels, representing individual classes, will be integrated into a hybrid modelling framework as shown in figure 2. Each input sequence will include the individual's assigned class, wearable-derived physiological signals, and pollution variables all spatially linked through GPS coordinates. These will be fed into a time-series transformer architecture structured as an encoder-decoder model. The transformer encoder will learn latent representations of long-range

dependencies across time, leveraging multi-head attention to capture complex interactions between variables as explained in figure 3 (Vaswani et al., 2017)[10]. This latent representation will serve as an identity map, associating pollution levels with corresponding physiological responses. The transformer decoder will then use this representation to forecast future states, effectively predicting the body's response to environmental changes across a prediction horizon. To improve long-range forecasting performance, the model will recursively uses its own predictions as inputs for subsequent time steps, promoting robustness and maintaining low mean squared error (MSE) over longer time spans.

Once general trends have been captured by the global model, a fine-tuning phase will adapt the model to individual-level data. This personalised calibration will enhance the model's ability to deliver real-time alerts, enabling proactive intervention by predicting individual physiological responses under anticipated pollution conditions.

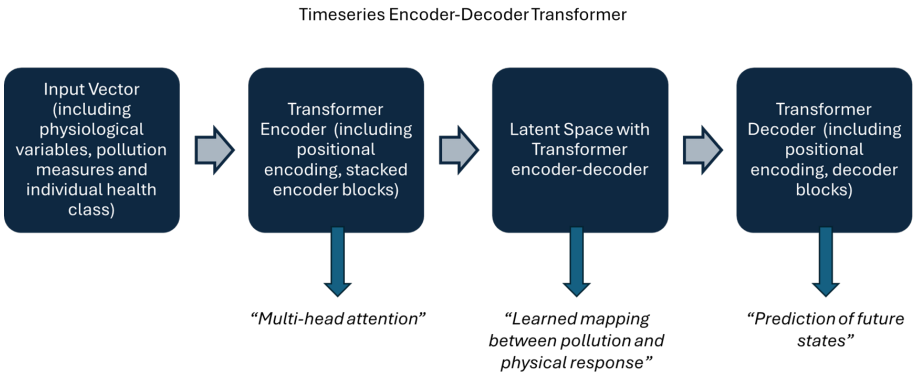

Figure 3: Time-series Transformer architecture mapping input features to a latent space

We evaluated limitations by training our encoder–decoder transformer on a synthetic sine-wave to forecast 100 steps ahead. Encoder–decoder blocks and positional encodings were essential for time awareness with near-zero Mean Absolute Error. However, this performance reflects the sine wave's simplicity, and forecast error grows with horizon length, which should be quantified via further analysis.

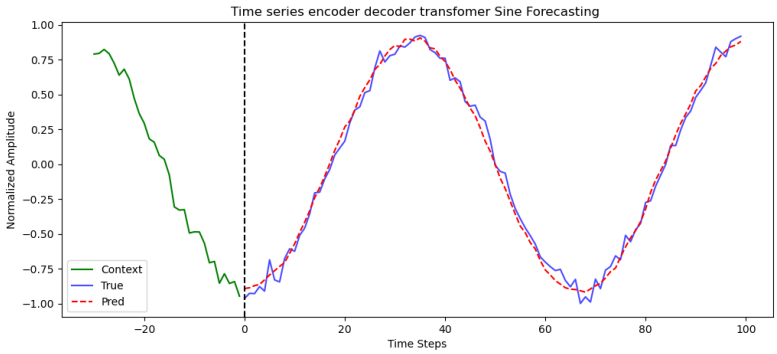

Figure 4: Time-series encoder-decoder predicting next step for sine wave

5. Timeline

Weekly Timeline

Main Task	Subtask	1 w	2 w	3 w	4 w	5 w	6 w	7 w	8 w	9 w	10 w	11 w	12 w	13 w	14 w
Project plan	Literature Review (Personalised Healthcare)														
	Literature Review (Transformer)														
	Model and Objective clarification														
Data-preprocessing	Data ingestion														
	EDA, Imputation & Outliers														
	Classification														
	Data loading and HPC														
Model Design	Understanding transformer model needs														
	Training and testing general model														
	Training and testing personal model														
Report writing															
Additional (optional)	Extending to clinical measures														

Figure 5: *Project timeline outlining key phases, including project plan, data preprocessing, model development, and report writing, with optional extension to clinical measures.*

The project is structured around four key phases, beginning with the definition of the research objectives and an in-depth literature review, as explained in figure 5. The subsequent phase involves the development of a data preprocessing pipeline to align the datasets and identify the most relevant features influencing physiological response.

Insights gained from both the literature and initial data analysis will guide the design of the model architecture. A time-series encoder-decoder transformer will be implemented to capture temporal dependencies and learn latent representations that form an identity map between pollution exposure and physiological response. These representations will be used to predict future physiological states under varying pollution conditions.

Several potential challenges have been identified that may impact the timeline. First, missing or corrupted data may compromise model reliability. As a mitigation strategy, any feature with more than 70% missing values will be excluded, as

such sparsity limits the ability to distinguish between true outliers and artefacts. Second, model interpretability poses a notable limitation. To address this, SHAP will be used to quantify the contribution of each input feature and improve the transparency of the transformer model predictions (Lundberg & Lee, 2017)[9].

If, after the training of the global model and subsequent fine-tuning at the individual level, performance remains unsatisfactory, additional clinical data will be integrated. These include lung function metrics, cardiovascular inflammation markers, and stress-related indicators. While these data may be temporally constrained, they offer additional context that could enhance the model's predictive power.

The ultimate objective is to develop a robust and interpretable framework capable of delivering personalised forecasts of physiological responses to pollution exposure, contributing to both individual-level health monitoring and broader public health strategies.

6. References

1. World Health Organization Overview (2025) Available at: https://www.who.int/health-topics/air-pollution#tab=tab_1 (09 June 2025).
2. Roos, L.G. and Slavich, G.M. (2023) 'Wearable Technologies for health research: Opportunities, limitations, and practical and conceptual considerations', Brain, Behavior, and Immunity, 113, pp. 444–452. doi:10.1016/j.bbi.2023.08.008.
3. He, Q. and Ji, X. (James) (2021) 'The labor productivity consequences of exposure to particulate matters: Evidence from a Chinese National Panel Survey', International Journal of Environmental Research and Public Health, 18(23), p. 12859. doi:10.3390/ijerph182312859.
4. Bernasconi, S., Angelucci, A. and Aliverti, A. (2022) 'A scoping review on wearable devices for environmental monitoring and their application for Health and Wellness', Sensors, 22(16), p. 5994. doi:10.3390/s22165994.
5. Hu, K. et al. (2014) 'Personalising pollution exposure estimates using wearable activity sensors', 2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), pp. 1–6. doi:10.1109/issnip.2014.6827617.
6. Verma, A., Ranga, V. and Vishwakarma, D.K. (2024) 'Breath-net: A novel deep learning framework for no2 prediction using bi-directional encoder with Transformer', Environmental Monitoring and Assessment, 196(4). doi:10.1007/s10661-024-12455-y.
7. Atzeni, M. et al. (2025) 'A machine learning framework for short-term prediction of chronic obstructive pulmonary disease exacerbations using

personal air quality monitors and lifestyle data', Scientific Reports, 15(1). doi:10.1038/s41598-024-85089-2.

8. Li, Y. et al. (2020) 'Behrt: Transformer for Electronic Health Records', Scientific Reports, 10(1). doi:10.1038/s41598-020-62922-y.

9. Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions (NeurIPS'17). In Neural Information Processing Systems (NeurIPS'17), 17212–17223.

10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I., 2017. Attention is all you need. In Advances in Neural Information Processing Systems (NIPS 2017).

11. AlNuaimi, A. F. A. H. and AlBaldawi, T. H. K. (2024) 'An overview of machine learning classification techniques', *BIO Web of Conferences*, 97, 00133. doi:10.1051/bioconf/20249700133.

12. Royston, P. and White, I. (2011) 'Multiple imputation by chained equations (MICE): Implementation instata', Journal of Statistical Software, 45(4). doi:10.18637/jss.v045.i04.