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

Independent Research Project
Final Report

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

August 29, 2025

Contents

Chapter 1

Abstract

Air pollution remains a major global health and environmental concern, contributing to an estimated seven million deaths annually because of the combined effects of out- door and household exposure (WHO, 2025)[1]. 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].

While recent research improved pollution forecast, there are gaps in understanding how these predictions affect individuals' health. 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 population model to capture general population temporal trends, and a personalised one specialised on individual characteristics. The population model is a transformer variational encoder–decoder, where the encoder captures long-term dependencies and the variational latent space supports in producing realistic decoding forecasts. The personalised model then adapts the population trends to unseen individual data. Our findings show not only that forecasting future hourly physiological states is feasible but it also suggest that different patients are more or less reactive to pollution. More sensitive ones can increase their breath rate to up to 10% when we increase pollution levels by six times.

Chapter 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. Cardiovascular and respiratory diseases as well as lung cancer are just some of the negative effects that the National Health System attributes to pollution (GovUK, *Health matters: Air pollution* 2018) [4].

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) [5]. Pollutants are categorised as either primary, such as PM, CO, and NO, or secondary, formed through chemical reactions like O_3 and NO_2, often found far from their original sources. Even though this study focused primarily on criteria pollutants such as PM_{10} and $PM_{2.5}$, due to their severe health risks (Bernasconi, Angelucci, & Aliverti, 2022) [5], other pollutants like NO, NO_2, O_3, SO_2, CO were examined for completeness.

Historically, air quality has been monitored using fixed-location stations, providing aggregated environmental data at a city or regional level. While useful for assessing general air quality trends, this approach presents two major limitations.

Firstly, fixed locations overlook personal exposure to pollution. These static measurements fail to capture the highly personalised nature of pollution exposure, which varies significantly depending on a person's location, mobility patterns, and daily activities (Hu et al., 2014) [6]. As a result, population-level estimates often obscure the true, specific impact of air pollution on human health (Hu et al., 2014) [6]. For instance, walking, jogging, or commuting through high-traffic areas can expose individuals to different pollution levels even within the same location (Hu et al., 2014) [6]. The same pollutant concentration may cause varying physiological responses across individuals,

depending on factors such as health status, age, pre-existing respiratory conditions, and lifestyle (Hu et al., 2014) [6].

Secondly, individuals' inhalation rates drastically change the amount of pollutants they absorb over a fixed period of time. This is critical because including the amount of air and its relative pollution level helps define the actual pollution inhalation rate of patients rather than relying only on general pollution measures (Lu & Fang, 2014) [7].

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_2 levels in India with high accuracy (Verma et al., 2024) [8]. However, it does not explore how these pollution fluctuations affect individual or population health, making it less useful for policymaking or preventative healthcare.

In contrast, Atseni et al. (2025) [9] developed a machine learning pipeline for short-term respiratory disease prediction. Their work underscores the importance of categorising 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)[10].

The value that this research is trying to add is not only to use wearable devices data to quantify the physiological response to pollution from individuals, but also to understand how much of the pollution surrounding the individual was practically inhaled with changing environmental conditions.

With this in mind, a transformer-based time series deep learning variational (GAN) encoder decoder architecture has been developed. The first objective is to leverage the transformer's strength in modelling long-range temporal dependencies to identify population-level trends. Thereby enabling the construction of an identity map, a variational latent space, that relates 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.

Chapter 3

Dataset

3.1 INHALE Dataset

The first dataset is provided by the INHALE project (Imperial College London, *Inhale*)[11] 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.

The INHALE dataset was the result of merging respiratory and pollution patients' data at different time steps. The datasets, connected through a common patient id identifiers , were useful to understand how different levels of pollution, including primary and secondary would affect individuals. Breath rate averaged over minutes or its standard deviation were useful measures to understand patient's reactivity to pollution.

The INHALE dataset temporal coverage was extremely limited. The data was collected during two distinct two-week periods in summer and winter, with non-continuous time steps. After some preprocessing , and specifically because of transformers time series needs, data was split into 1-hour long blocks where any shorter block was filtered out and not considered. This was useful to give the model enough data to understand how patterns would evolve in the long-run.

Dates and time fields also needed to be processed. Indeed, left as is, neural networks tend to treat each timestamp as a distinct category, focusing more on nearby ones while underestimating longer range relationships.

To fix this, we encoded time as cyclic signals. We extracted hour, day of week, and day of year, then mapped them onto sine and cosine functions so that times like 23:00 and 00:00 are treated as neighbours rather than distant values. This preserves natural cycles and helps the model learn daily, weekly, and seasonal patterns more effectively (Invidia Developer Forum, 2022)[12].

3.2 OpenWeather Dataset

The second dataset, coming from OpenWeather API, provides real-time air pollution levels, including key pollutants such as PM2.5, PM10 NO2 etc (*Current weather and forecast - openweathermap* 2025)[13]. Even though PM2.5 and PM10 were identified as the pollutants with the greatest consequences for individuals' health, other pollutants like no, no2 , o3 and so2 were considered for a more complete approach, as per figure below. Indeed, considering different pollutants is a good way to summarise not only inhalation rates when commuting, but also any primary or secondary agents inhaled when spending time indoor and outdoor (Jonidi Jafari et al., 2021)[14].

caption

Pollutant	Sources
PM	Transport (including exhaust fumes and tire & brake wear); combustion; industrial processes; construction & demolition; wind erosion
NO_2	Combustion processes (heating, power generation, and engines in vehicles & ships)
SO_2	Use of sulfur-containing fossil fuels for domestic heating, power generation, and vehicles
CO	Transport (especially petrol engines); combustion; industry
O_3	Photochemical reaction of NO_x (from industry & vehicles) with VOCs (from automobiles, industry, solvents)

Figure 1: *Pollutants by source – useful to understand impact on patients*

The Openweather dataset is geolocated using longitude and latitude coordinates, offering hourly pollution data at a spatial resolution of approximately 200 metres. When matched with the individual respiratory and pollution's data coming from the Inhale dataset, it enables a spatially aware analysis of personal exposure to air pollution, allowing for the integration of environmental data with individual physiological responses. The hourly and 200-metre resolution is one of the key limitations of this study, because it does not help us capturing highly specific variations in pollution across different times of the day.

3.3 Merged Dataset

The Inhale Dataset, coming from INHALE was therefore merged with the OpenWeather dataset based on longitude, latitude and timestamp. This was key, since it gives us a map of how individuals moved across London, including different levels of indoor and outdoor pollution as well as their physiological responses to it.

Pollution levels together with individuals' physiological responses from patients across time and inhalation rates were the backbones of our dataset resulting in a highly accurate and specialised model.

Chapter 4

Methodology

4.1 EDA and Feature Correlation

After merging the dataset and deploying data preprocessing techniques, some exploratory data analysis was implemented so that the transformer model could later be trained.

Outliers removal was an essential first step. Given the data collection process and the research objectives, outliers were handled manually rather than with standard inter quartile range methods. This choice reflects the importance of maintaining real spikes, f.i. a sudden increase in breath rate that may indicate reactions to pollution. At the same time, implausible values were excluded, for example temperatures of 56 °C in March or pollution levels high enough to be immediately lethal. These unrealistic records were removed from the dataset.

After removing outliers, the data was normalised. Standard normalisation was necessary because the features had very different scales, and without adjustment, variables with larger magnitudes would have disproportionately influenced the model.

In addition, 12 patients were completely excluded. This is because their breath rate average feature was missing around 30/40% of total value, leading to imputation as the only viable option.

After cleaning and processing the data, a correlation heatmap, as shown in figure 2, was generated to examine relationships between features. As expected, breath rate and physical activity features have some correlation, greater levels of physical activity leads to more frequent breaths and so greater breathing rate. Strong correlations also appeared among pollution measures, consistent with the fact that polluted environments typically contain multiple pollutants. There is no evident correlation between physiological measures and pollution measures from the correlation map.

A Pearson correlation heatmap, as the one below, is mainly used to define linear relationships, while the relationship between patients' physiological data and pollution is non-linear. Only by looking at multi-features over different time steps we could

identify any relationship between the two.

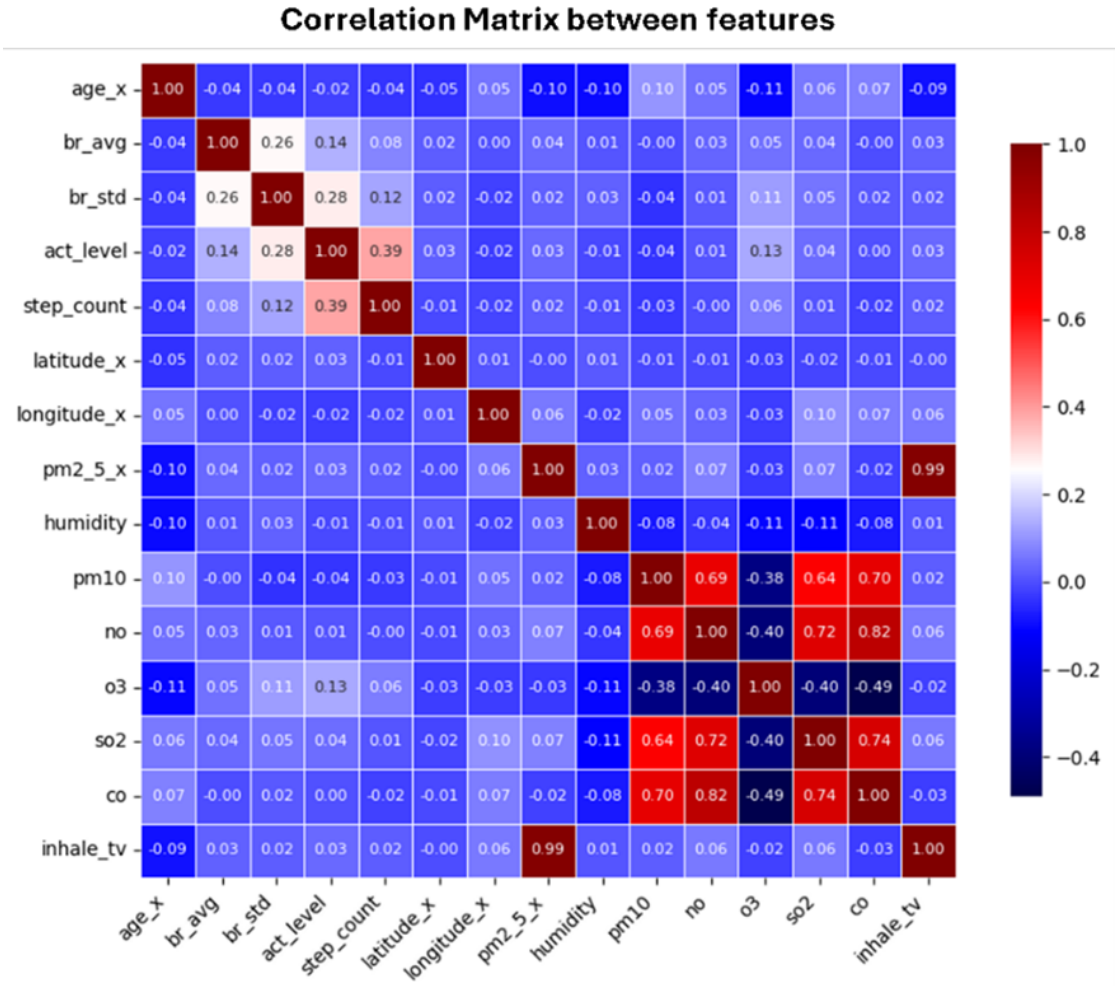

Figure 2: Correlation Matrix between different features within pollution measures and within physiological measures. No strong cross-category correlation

4.2 Data Loading (Sliding Window)

We randomly took 30 of the 44 patients for training and 13 for testing. This is key because the model needs both healthy and asthmatic patients. Keeping the original sequence order would have biased the results.

A custom data loader was defined and applied to both training and test data. The first step was to determine how many windows to create using the formula:

Number of windows $= (numberofsamples - windowsize - forecaststeps)\ //\ step$

This shows that the number of windows depends on the dataset size, reduced by the number of past steps and future steps, then adjusted by the step size to optionally skip rows for memory efficiency. Each window was then split into an input and a target. The input represents the historical data used for training while the target is the future data the model must predict. With this setup, the model iteratively learns by processing each input block and predicting its corresponding target.

4.3 Model Architecture

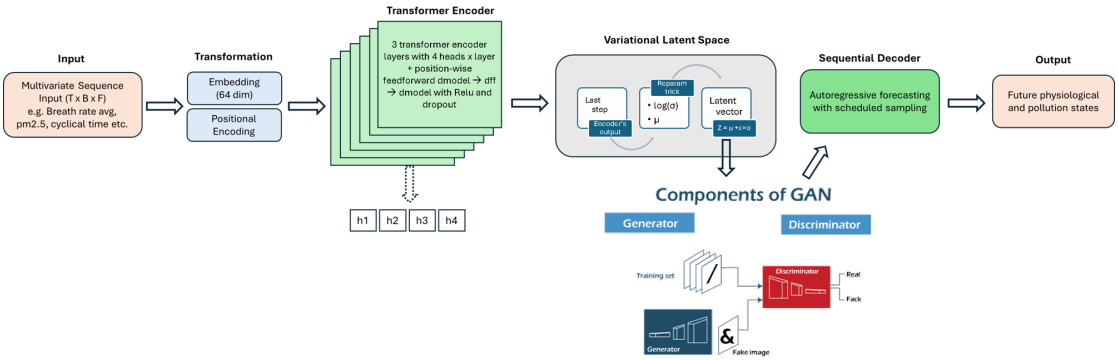

Figure 3: Transformer Vae Gan (Thatipalli, 2023)[15] Encoder Decoder design structure

4.3.1 Transformer Encoder

The model takes in multivariate sequences of physiological and environmental data, such as breathing rate, activity, and pollution measures across different time steps. The input is structured as a 3d tensor of sequence length, batch size and number of input features. Each time step is embedded into a higher-dimensional vector (64 dimensions but customisable) and with added sinusoidal positional encodings, so that temporal order between different time steps is preserved.

Once embedded and positionally encoded, the sequence is then passed through a stack of six transformer encoder layers.

Each encoder layer has two main parts: multi-head self-attention and a feedforward network. Self-attention lets each time step relates to the whole sequence, capturing both immediate and delayed responses (Casolaro et al., 2023)[16] . This is crucial for physiological data, where reactions to pollution may build up or being imminent. The model uses eight attention heads, with some attending to short-term fluctuations and others to longer-term patterns, reflecting how individuals respond to both recent and past exposures(Vaswani et al., 2017)[17].

After self-attention, the outputs pass through a two-layer feedforward network with a ReLU activation in between. The first layer expands the representation from the embedding size (d_model) to a larger hidden size (dim_feedforward), giving the model more capacity to learn. ReLU adds non-linearity, and dropout reduces overfitting by randomly dropping activations. The second layer then projects the representation back to the original embedding size.

4.3.2 Variational Latent Space

After the input sequence passes through the Transformer encoders, the final hidden state is taken as a summary of the input sequence. This then gets projected into two vectors: a mean (μ) and a log-variance ($\log(\sigma^2)$), which together define a multivariate Gaussian distribution. Instead of mapping inputs to a single latent vector, we use a variational approach that will sample from this distribution(Mao et al., 2020)[18]. The standard deviation is computed as:

$$\sigma = \exp\Big(0.5 \times \log(\sigma^2)\Big)$$

To keep the sampling step differentiable, we apply the reparameterisation trick:

$$s = \mu + \epsilon \times \sigma, \quad \epsilon \sim \mathcal{N}(0, I)$$

The sampled latent vector s acts as a bottleneck representation, keeping the key temporal information needed for forecasting. Combining this variational latent space with the transformer encoder allows the model to capture both temporal dependencies and uncertainty, making it powerful for physiological time-series prediction.

4.3.3 Training

The model is trained with an adversarial autoencoding framework designed for time-series forecasting in mind. Training runs in two distinct phases.

In the first phase, a discriminator learns to distinguish between latent vectors coming from a standard normal distribution (s_real) and those generated by the encoder (s_fake). This pushes the encoder to align its latent space with the Gaussian distribution. For each forward pass, a new latent vector is sampled from the encoder's distribution and evaluated by the discriminator. As a result, an adversarial loss is added to the total loss. In the second phase, the decoder is trained to forecast the next time step block of the sequence, with the mean squared error as the result between the predicted and actual values.

To improve robustness, forecast is implemented autoregressively and with scheduled sampling. The model indeed, gradually shifts from using true input(1 hour) to using its own predictions when forecasting next steps (12 hours in blocks of 60mins).

Together, these steps combine accurate forecasting with latent space regularisation, improving both predictive performance and generalisation.

Chapter 5

Results and Discussion

We start discussing model results by focusing on general trends from the 43 patients trained model. We begin by explaining why forecasting was the main focus rather than reconstructing the input and we will then analyse clustered reactions to pollution. We will delve deeper to discover whether asthmatic individuals show stronger or weaker responses than healthy ones. Pollution levels will then be perturbed to observe changes in physiological measures.

After discussing general model results, we focused on generalisation by fine-tuning the general model on data from one unseen individual. Finally, we propose a threshold-based alert system that forecasts an individual's physiological responses to pollution and triggers warnings when critical thresholds are expected to exceed.

5.1 Prediction

Using a multimodal time series model combining pollution concentrations, inhalation rate, and physiological signals, we predicted next-hour physiological and pollution measures aggregated across all subjects.

As shown in figure 4, the model is able to track closely both physiological measures such as breath average and pollution levels, with large spikes being the hardest to capture. Two design choices are key to this performance. The former is the ability of transformers to capture temporal dependencies, where past inputs affect both present and future timesteps. The latter comes from adversarial training, where generator and discriminator are working together to make the predictions as realistic as possible.

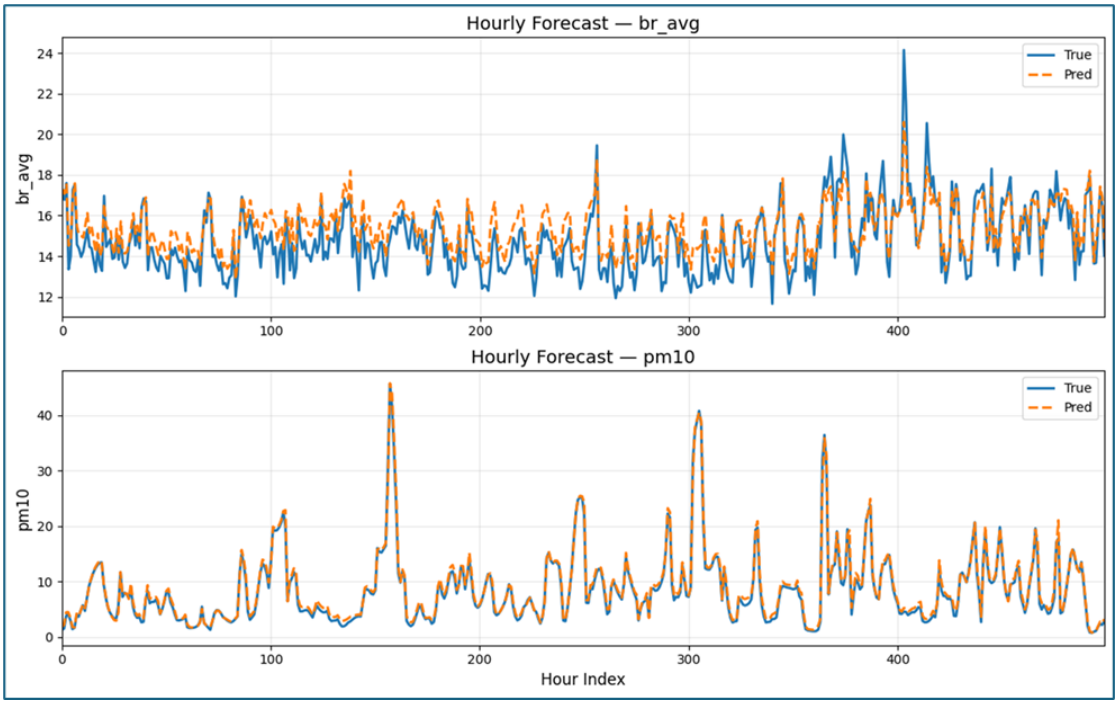

Figure 4: Prediction of breathing rate average (br_avg) and pollution (PM_{10}) on an hourly basis, aggregated across all patients

Also, we decided to focus on forecasting rather than reconstructing inputs to anticipate how an individual's physiology responds to pollution. The learned latent space captures person-specific dynamics, making forecasts valuable to provide individuals with clear insights and actionable support.

5.2 Cluster Reactions

After showing that forecasting next step was feasible and accurate, we analysed how patients reacted to pollution and whether subgroups had stronger or weaker responses. For each patient, sliding windows were passed through the model to extract latent vectors (z), which were then averaged to produce a single embedding per individual. These embeddings were clustered using K-Means to identify groups with similar response patterns, as shown in figure 5.

Most individuals clustered closely together, reflecting the Gaussian distribution learned by the adversarial model, while a few formed distinct groups, suggesting a few individuals respond differently to pollution.

For a better understanding of group-specific responses to pollution, we examined the distributions and medians of pollution and physiological measures across the seven

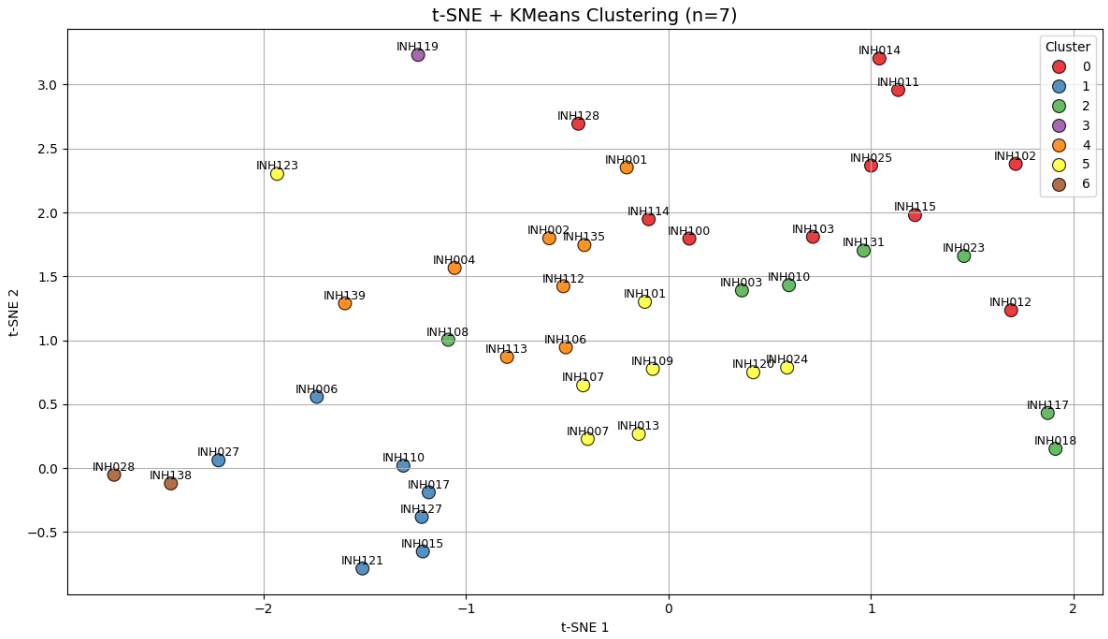

Figure 5: K-cluster for all patients in seven distinct groups including healthy and asthmatic individuals

clusters.

When looking at figure 6, we can straight away see that cluster 3 was characterised by the highest levels of $\mathrm{PM}_{2.5}$ and PM_{10} levels and shows one of the largest increases in average breathing rate. However, it does not result in the strongest one. This is likely because activity levels in this group are relatively low. Cluster 4 illustrates this relationship really well: despite slightly lower $\mathrm{PM}_{2.5}$ and PM_{10}, higher activity levels are associated with the strongest breathing-rate response. In contrast, Cluster 0 has modest $\mathrm{PM}_{2.5}$ but elevated O_3 and SO_2 relative to other clusters, consistent with vehicles and indoor pollution influences. Individuals in cluster 0 appear more affected by these pollutant mixtures.

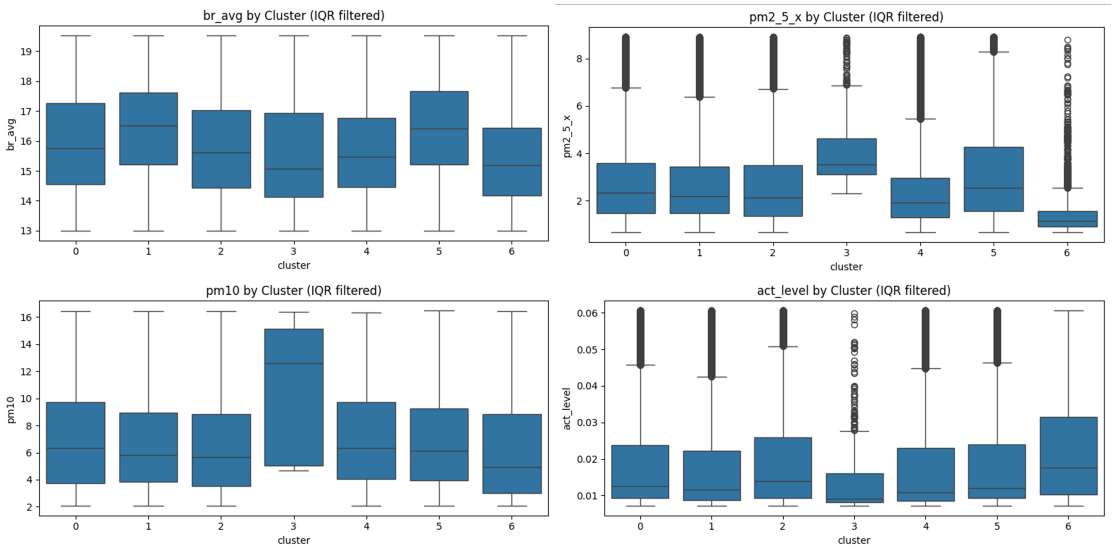

Figure 6: Distribution of breath rate average, pm2.5, activity level and pm10 by seven clusters

5.3 Healthy vs Asthmatic Individuals

Besides clustering individuals based on their response to pollutants, we also decided to analyse whether healthier individuals respond to pollution any differently than asthmatic ones. We divided our dataset in two categories, where asthmatic individuals were 19 while healthy ones were 24. In addition, the healthy patients group have inherently more data which involves the model being able to forecast, reconstruct and understand healthy individuals better.

Although the literature suggests that asthmatic individuals tend to react more strongly to pollution than healthy ones (Kim et al., 2013)[19], our study does not show that. This was expected for a few different reasons. Healthy individuals were exposed to higher levels of $\mathrm{PM}_{2.5}$ than asthmatic. This, together with higher activity levels induced a more intense level of inhalation rate across time for healthy patients.

In practice, greater pollution exposure, higher activity levels, driven by healthy individuals commuting more than asthmatics, resulted in a more critical breathing-rate response among healthy participants, as illustrated in the figure 7 below.

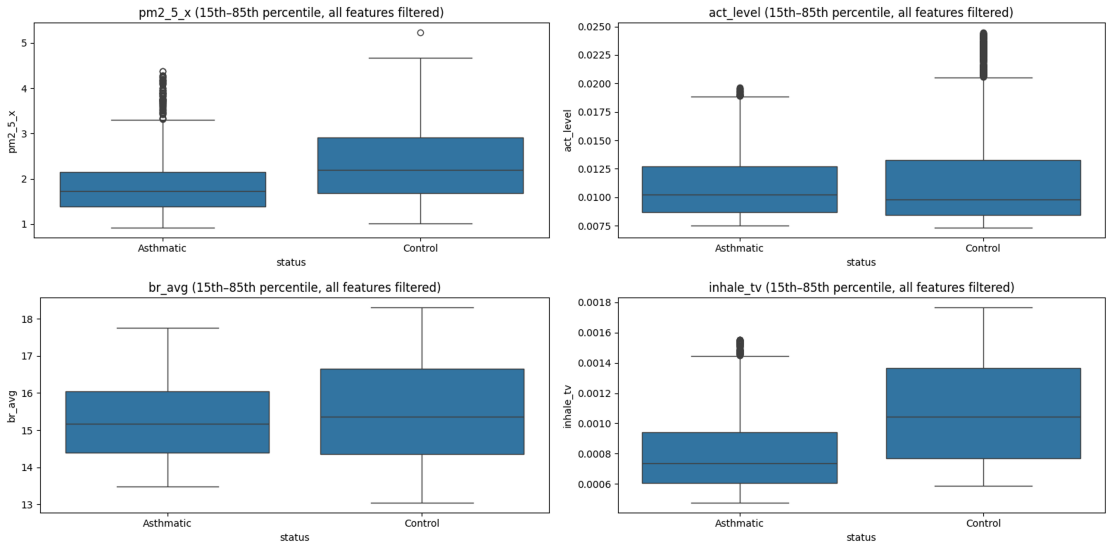

Figure 7: Distribution of pm2.5, activity, inhalation rate and breath rate for healthy (control) vs asthmatic patients

5.4 Perturbation Experiments

After comparing healthy and asthmatic groups, we assessed cluster-level reactivity to pollution using a perturbation experiment. Pollution inputs were artificially scaled (1×, 2×, 4×, 6×) while all other variables including activity, temperature, humidity, inhalation volume, and time encodings remained fixed. The perturbed data was windowed, normalised and passed through the model to forecast one step ahead. For each scenario, we examined predicted physiological outcomes (e.g., average breathing rate) across clusters. Results were pooled within clusters and visualised with boxplots, trimming the 10–90% range to reduce the influence of extreme values.

From picture 8 below, we noticed that greater levels of pollution lead to higher volatility within breathing rate, meaning more unstable breathing. This phenomenon is consistent across all clusters. When looking though at breath rate average per cluster, we can see how specific clusters, such as 1 and 3 are more reactive to pollution while the others do not react as strongly.

As shown above, comparing the baseline scenario with the most extreme case (x6) reveals distinct cluster-specific responses. Clusters 1, 3, and 6 show the strongest reactivity, with breath rate rising by up to 10%. In contrast, Clusters 0, 2, 4, and 5 display only mild changes in average breath rate but reflect more erratic standard deviations, indicating spikier breathing patterns.

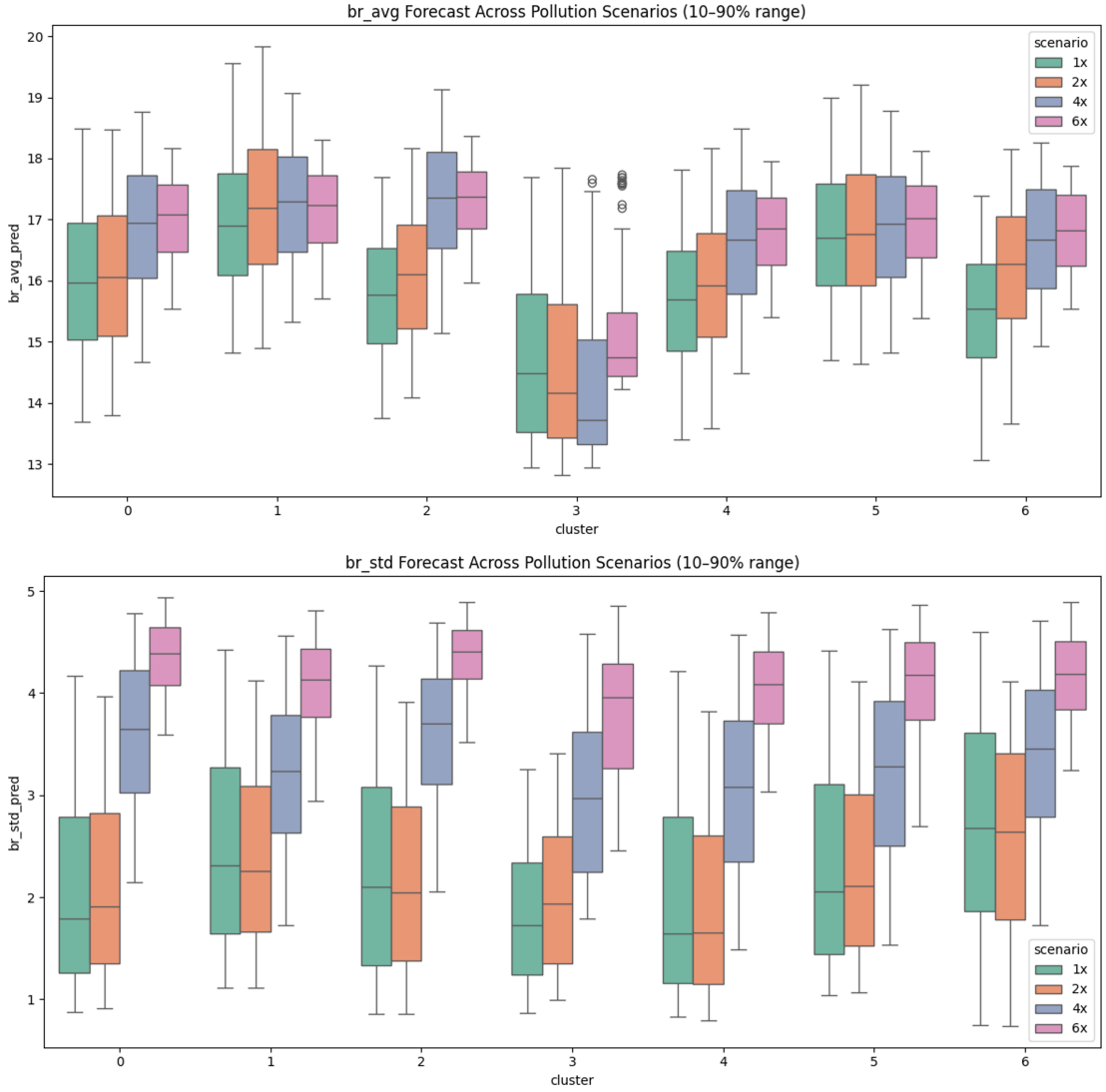

Figure 8: Perturbation of pollution measures by clusters (1 = baseline, 6 times the baseline)

5.5 Generalisation

After forecasting future steps for all individuals and identifying which clusters were more or less reactive to pollution, we tested whether the general model could generalise to unseen individuals.

As shown in figure 9, the model successfully predicted hourly physiological responses for new patients. This shows that the model captured general trends from the training patients while also adapting to new features, patterns, and edge cases in unseen

data. Proving its ability to generalise to entirely new dynamics.

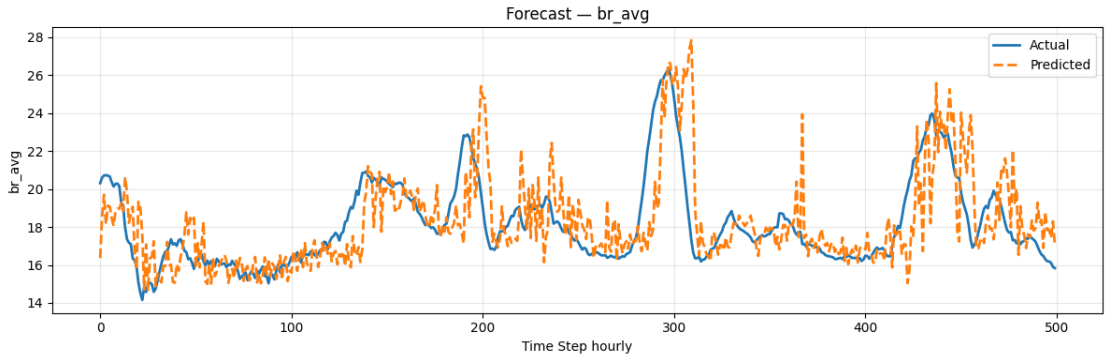

Figure 9: Forecast of unseen data hourly for both breath rate average and standard deviation for an individual

5.6 Threshold-Based Alerts

The aim of the alert system is to detect periods of extreme physiological change caused by pollution exposure. Firstly, a baseline profile for each individual is established. Then deviations from that baseline with new time steps are computed, flagging high-risk events in case of both pollution and physiological deviations being abnormally high.

The baseline is created by passing all sliding windows coming from the validation dataset through the model and extracting the first forecasted timestep per feature. These forecasts are inverse transformed and averaged to produce a reference vector representing the individual's expected physiological state under normal conditions.

New forecasts are then compared to this baseline, generating a deviation score that quantifies how far the physiology has shifted from normal behaviour. Repeating this process creates a continuous deviation series alongside pollution values at each timestep.

Risk, instead was defined using percentile thresholds. Pollution values above the 75th percentile were treated as high exposure, while those below the 25th percentile were treated as low exposure. Since these thresholds are calculated from each individual's own data, they are customised to different patients baseline vector, adapting automatically to different individuals.

An alert is then triggered only when both conditions are met: pollution exceeds its 75th percentile and the physiological deviation also exceeds its own threshold. These events, highlighted as red dots in the final visualisation, indicate moments when environmental stress and physiological stress coincide. Indeed, because the alert system gets triggered only when physiological deviation and pollution measure are both high, false alarms are skipped. Situations where only pollution is high but there is no reaction to it from the chosen patient, or physiological measures deviation is abnormal but not driven

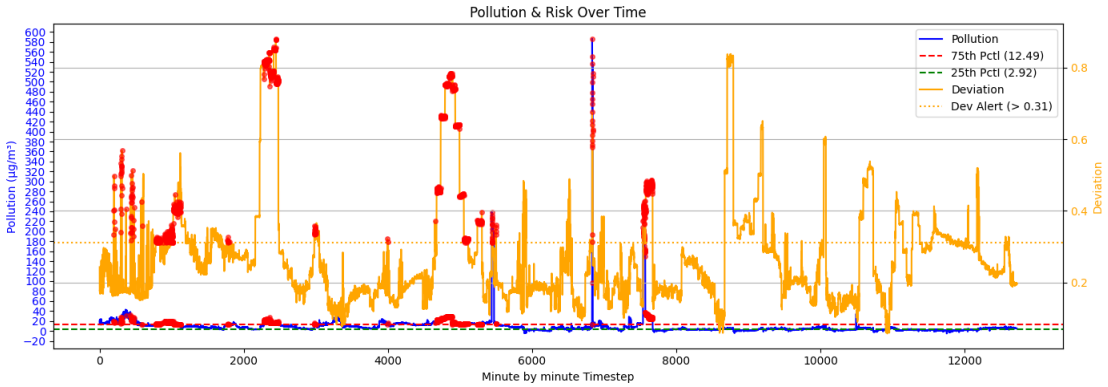

Figure 10: Hourly Threshold alerting system with pollution level in blue, deviation from standard in yellow and red points potential alerts

Chapter 6

Limitations

Even though the overall number of patients included in the study was representative enough, there was an imbalance between them being healthy or asthmatic. Healthy individuals contributed more to the data, resulting in a model better at capturing healthy responses vs asthmatic ones.

The entire dataset was collected in London. This spatial restriction is a key limitation because it weakens generalisability. Pollutants, weather conditions and commuting patterns varies across different regions, so the model won't generalise well with other cities or countries.

The physiological measures were also limited. Breath rate and its evolution over time is not enough to quantify the full effects of pollution on individuals. Additional signals, such as heart rate, stress levels would provide a more complete map of how individuals react to pollution in the short and long-term.

The OpenWeather dataset also brings some additional limitations. Its hourly temporal resolution, preventing minute by minute analysis, and its 200metres spatial resolution is not able to capture micro-environmental variations (roadside vs indoor exposure). These limitations are especially relevant for commuting or other short-term activities where exposure changes quickly.

Finally, the model architecture also has some limitations. The Transformer–VAE–GAN architecture, while powerful in capturing nonlinear temporal dependencies, operates as a black box. Its internal representations make it difficult to attribute predictions to specific pollutants or physiological drivers reducing the model's suitability for clinical applications.

Chapter 7

Conclusion

After conducting exploratory data analysis on the data, implementing data preprocessing techniques, a clean set of data was fed to the model. Then, a Transformer Vae Gan was trained with the aim to forecast next time steps, including both physiological and pollution measures.

After the model has been developed and results have been analysed, we clearly identified that the model was precise enough to forecast next steps both in terms of general trends but also fine-tuned to unseen individuals. In addition, after applying a k-clustering technique to individuals' latent space, we can see how different clusters react differently to similar levels of pollutions. Some clusters were more reactive to pollution, with spikes reaching 11% (when pollution is increased 6 times) while other clusters having a mild reaction to perturbation.

In addition, a threshold alert system was developed to operationalise the model. Indeed, being able to alert individuals on how they will physically react when experiencing similar pollution levels could be a great aid especially for those clusters that are very sensitive to pollution. Another activity, or a different way to commute could be suggested to reduce overall pollution exposure and inhalation rates.

Appendix A

Appendix

Tables that could not be included because exceeding the 10 pictures limits but still valid

General Model MSE and MAE by features	Mean Square Error	Mean Absolute Error
br_avg	0.379	0.458
br_std	0.251	0.360
act_level	0.000	0.005
step_count	9.273	1.240
pm2_5_x	22.375	3.089
temperature	0.508	0.544
humidity	0.720	0.668
hour_sin	0.004	0.038
hour_cos	0.006	0.049
dow_sin	0.003	0.024
dow_cos	0.003	0.022
yearly_sin	0.004	0.021
yearly_cos	0.003	0.042
lat_round	0.013	0.013
lon_round	0.003	0.024
pm10	1.190	0.613
no	23.724	1.677
no2	5.234	2.003
o3	20.344	3.725
so2	2.858	1.307
co	214.428	8.171
inhale_tv	0.000	0.001

Table A.1: General Model Mean Squared Error and Mean Absolute Error

Individual Model MSE and MAE by features	Mean Square Error	Mean Absolute Error
br_avg	1.641	0.909
br_std	0.939	0.653
act_level	0.000	0.007
step_count	26.015	1.644
pm2_5_x	224.903	4.198
temperature	1.057	0.567
humidity	1.456	0.584
hour_sin	0.017	0.055
hour_cos	0.017	0.050
dow_sin	0.004	0.033
dow_cos	0.005	0.025
yearly_sin	0.000	0.011
yearly_cos	0.002	0.015
lat_round	0.000	0.004
lon_round	0.000	0.018
pm10	2.613	0.697
no	59.641	3.891
no2	7.544	1.265
o3	21.391	1.613
so2	5.020	1.059
co	385.518	7.911
inhale_tv	0.000	0.001

Table A.2: Individual forecast Mean Squared Error and Mean Absolute Error

References

1. World Health Organisation (2025). *Overview*. Available at: https://www.who.int/health-topics/air-pollution#tab=tab_1

2. Roos, L.G. & Slavich, G.M. (2023). Wearable Technologies for health research: Opportunities, limitations, and practical and conceptual considerations. *Brain, Behavior, and Immunity*, 113, 444–452. doi:10.1016/j.bbi.2023.08.008

3. He, Q. & Ji, X. (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), 12859. doi:10.3390/ijerph182312859

4. UK Government (2018). *Health matters: Air pollution*. Available at: https://www.gov.uk/government/publications/health-matters-air-pollution/health-matters-air-pollution

5. Bernasconi, S., Angelucci, A., & Aliverti, A. (2022). A scoping review on wearable devices for environmental monitoring and their application for health and wellness. *Sensors*, 22(16), 5994. doi:10.3390/s22165994

6. Hu, K. et al. (2014). Personalising pollution exposure estimates using wearable activity sensors. In *Proceedings of IEEE ISSNIP*. doi:10.1109/issnip.2014.6827617

7. Lu, Y. & Fang, T. (2014). Examining personal air pollution exposure, intake, and Health Danger Zone using time geography and 3D geovisualisation. *ISPRS International Journal of Geo-Information*, 4(1), 32–46. doi:10.3390/ijgi4010032

8. Verma, A., Ranga, V., & Vishwakarma, D.K. (2024). Breath-net: A novel deep learning framework for NO_2 prediction using bi-directional encoder with Transformer. *Environmental Monitoring and Assessment*, 196(4). doi:10.1007/s10661-024-12455-y

9. Atseni, M. et al. (2025). A machine learning framework for short-term prediction of chronic obstructive pulmonary disease exacerbations. *Scientific Reports*, 15(1). doi:10.1038/s41598-024-85089-2

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

11. Imperial College London (2024). *INHALE Project*. Available at: https://www.imperial.ac.uk/earth-science/research/research-projects/inhale/

12. NVIDIA Developer Forum (2022). Three approaches to encoding time information as features for ML Models. Available at: https://developer.nvidia.com/blog/three-approaches-to-encoding-time-information-as-f

13. OpenWeather (2025). *Current weather and forecast*. Available at: https://openweathermap.org/

14. Jonidi Jafari, A., Charkhloo, E., & Pasalari, H. (2021). Urban Air Pollution Control Policies and strategies: A systematic review. *Journal of Environmental Health Science and Engineering*, 19(2), 1911–1940. doi:10.1007/s40201-021-007444

15. Thatipalli, A. (2023). Generative Adversarial Networks (GANs): A simple explanation. Medium. Available at: https://ai.plainenglish.io/generative-adversar

16. Casolaro, A. et al. (2023). Deep learning for time series forecasting: Advances and open problems. *Information*, 14(11), 598. doi:10.3390/info14110598

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

18. Mao, X. & Li, Q. (2020). Generative Adversarial Networks (GANs). In *Generative Adversarial Networks: Architectures and Applications*.

19. Kim, K.-H., Jahan, S.A., & Kabir, E. (2013). A review on human health perspective of air pollution with respect to allergies and asthma. *Environment International*, 59, 41–52. doi:10.1016/j.envint.2013.05.007