Open Access

Advances in Environmental and Engineering Research

LIDSEN Publishing Inc.

Original Research

Estimating Inhaled Nitrogen Dioxide from the Human Biometric Response

Shisir Ruwali, Bharana Ashen Fernando, Shawhin Talebi, Lakitha Wijeratne, John Waczak, Prabuddha M. H. Dewage, David J. Lary *, John Sadler, Tatiana Lary, Matthew Lary, Adam Aker

Department of Physics, University of Texas at Dallas, 800 W Campbell Rd, Richardson TX 75080, USA; E-Mails: Shisir.Ruwali@UTDallas.edu; ashen.fernando@utdallas.edu; shawhintalebi@gmail.com; lhw150030@utdallas.edu; John.Waczak@UTDallas.edu; PrabuddhaMadusanka.HathurusingheDewage@UTDallas.edu; David.Lary@utdallas.edu; sadlercjack@gmail.com; tlary@me.com; MDL210001@utdallas.edu; Adam.Aker@UTDallas.edu

* **Correspondence:** David J. Lary; E-Mail: David.Lary@utdallas.edu

Academic Editor: Pallav Purohit

Adv Environ Eng Res
2024, volume 5, issue 2
doi:10.21926/aeer.2402011

Received: December 19, 2023
Accepted: April 23, 2024
Published: May 08, 2024

Abstract

Breathing clean air is crucial for maintaining good human health. The air we inhale can significantly impact our physical and mental well-being, influenced by parameters such as particulate matter and gases (e.g. carbon dioxide, carbon monoxide, and nitrogen dioxide). Building on previous research that explored the effects of particulate matter (PM) in specific environments, analyzed using biometric indicators and machine learning models; this work focuses on the effects and estimation of inhaled nitrogen dioxide (NO_2). This study involved a cyclist equipped with sensors to monitor various biometric parameters. In addition, an electric car following the cyclist measured the ambient NO_2 levels using an onboard sensor. A total of 329 biometric variables have been taken into account, of which 320 biometric variables are cognitive responses extracted using an electroencephalogram (EEG) and 9 biometric variables are physiological responses extracted using several sensors. Inhaled NO_2 levels are first estimated initially by making use of all 329 variables, then using 9 physiological responses and finally using only 6 of the 9 physiological responses. The study also uses a ranking method to pinpoint which biometric variables most significantly estimate inhaled NO_2 levels.

Furthermore, it investigates the linear and non-linear relationship between certain variables and inhaled NO_2. The general precision of the prediction for the data set was moderate, as indicated by the coefficient of determination (R^2) and the root mean square error (RMSE) between the true and estimated values of NO_2 to be 0.35 and 5.41 ppb, respectively, in the test set. A higher accuracy in the prediction of lower values of NO_2 levels was qualitatively observed using a scatter diagram and a Quantile-Quantile plot where the data were more plentiful. For more robust conclusions, additional data and refined machine learning models are necessary.

Keywords
NO_2; particulate matter; machine learning; biometrics

1. Introduction

Recent studies have shown that more than half of the world's population is exposed to increasing levels of air pollution and almost 99% of the global population breathes air that exceeds the air quality standards set by the World Health Organization [1, 2]. Common air pollutants include particulate matter, carbon dioxide, carbon monoxide, oxides of nitrogen, lead, sulfur dioxide, and ground-level ozone. These pollutants are known to be the cause of many adverse health-related effects such as airway inflammation, decreased cognitive performance, respiratory issues with increased cough and inflammation of the lungs, cardiovascular disease, and metabolic effects [3-9].

Nitrogen dioxide is one of the six air pollutants for which air quality standards have been established by the United States Environmental Protection Agency to reduce its level in the outdoor environment [10]. This air pollutant usually forms when coal, diesel, and other fossil fuels are burned at high temperatures. Increased levels of NO_2 have been associated with cardiovascular and respiratory mortality [11]. Indoor NO_2 levels are known to aggravate respiratory problems in children with asthma causing frequent cough and wheezing; exposure to NO_2 has been associated with adverse health effects related to the lungs, respiratory system and increased hospitalization cases [12-15].

The studies mentioned showing the adverse health effects of inhaling NO_2 in the human body have typically not examined very fine temporal and spatial scales. In this study, we introduce a novel approach where machine learning models are used to estimate inhaled NO_2 using biometric variables of a participant and to understand the effects of inhalation of ambient NO_2 on a small temporal scale (≈10 s) and a small spatial scale (≈2 m).

This work is an extension of a previous study [16] that used observed autonomic responses of the body characterized by biometric sensors to accurately estimate inhaled particle concentrations, in particular, PM_1 and $PM_{2.5}$, using machine learning models from biometric variables of a single participant, and studied the effects of inhaled particles with very high precision with the coefficient of determination (R^2) between true and estimated PM_1 values of 0.91 in an independent test set. The estimation of inhaled CO_2 concentrations using machine learning from biometric variables of participant autonomic responses was also found to be highly accurate, with R^2 between the true and estimated values of CO_2 being 0.98 in an independent test set [17]. In a different study, the use

of biometric variables to estimate $PM_{2.5}$ with data collected from multiple participants was also found to be highly accurate, with R^2 between the true and estimated values of $PM_{2.5}$ being 0.99 in an independent test set [18]. Other studies have used machine learning models to accurately estimate $PM_{2.5}$ from a series of more than 30 meteorological and environmental data such as aerosol optical depth (AOD), temperature, humidity, etc. [19].

In this study, we considered the concentrations of inhaled nitrogen dioxide in ambient air and characterized the autonomic response, so that we can use only the autonomic response to estimate the inhaled nitrogen dioxide concentration. As air pollution that includes NO_2 as a component is known to have many adverse health effects related to cognitive performance, cardiovascular disease, respiration problems, and inflammation of the airways, as mentioned previously, several biometric variables were measured to capture as many autonomous cognitive and physiological responses as possible that are caused by inhaling outdoor air holistically. Biometric variables measured and used in this study include skin temperature, heart rate, respiration rate, electrocardiogram (ECG), galvanic skin response (GSR), blood oxygen saturation (SpO_2), pupil diameter of the left eye, pupil diameter of the right eye, distance between pupils, and measurement of electrical activity across the surface of the brain using a 64 electrode electroencephalography (EEG). These measurements were made when a participant was riding a bicycle outdoors equipped with a biometric suite with an electric car behind equipped with multiple sensors to capture ambient NO_2, PM, carbon dioxide (CO_2) and nitric oxide (NO).

Two of the main objectives of this study include (i) testing whether the methodology used to estimate and understand the effects of inhaled PM_1 and $PM_{2.5}$ from biometric variables of a person using machine learning can be extended to NO_2. (ii) studying the effects of inhaled NO_2 on human autonomic responses. One of the significant parts of the study is that it not only studies the effect of NO_2 on the human body, but also studies the relationship between the biometric variables and tests if they are mutually related to each other linearly or nonlinearly. We also use Occam's razor principle to test if a simpler model consisting of a smaller set of biometric variables can be used to produce similar or even better results.

2. Materials and Methods

The methodology implemented in this study includes two key parts (a) simultaneously measuring the biometric variables of a participant cycling a bicycle wearing a comprehensive biometric suite and a reference sensor measuring the ambient NO_2 concentration and (b) using machine learning models to estimate the inhaled NO_2 using the measured biometric responses of the participant.

A complete description of the procedure of data collection is given in the previous work [16] while a brief description is given below.

2.1 Experimental Paradigm and Holistic Sensing

The experimental suite of sensors is shown in Figure 1. Figure 1a shows a photograph of the participant wearing the biometric suite followed by an electric car equipped with the reference sensor that measures ambient NO_2. This biometric suite consisting of an array of devices measures the biometric variables (or features or predictor variables or autonomous cognitive and physiological responses) in the participant, such as: EEG, SpO_2, heart rate, respiration rate, ECG, GSR,

skin temperature, pupil diameter of the left eye, pupil diameter of right eye and distance between the pupils.

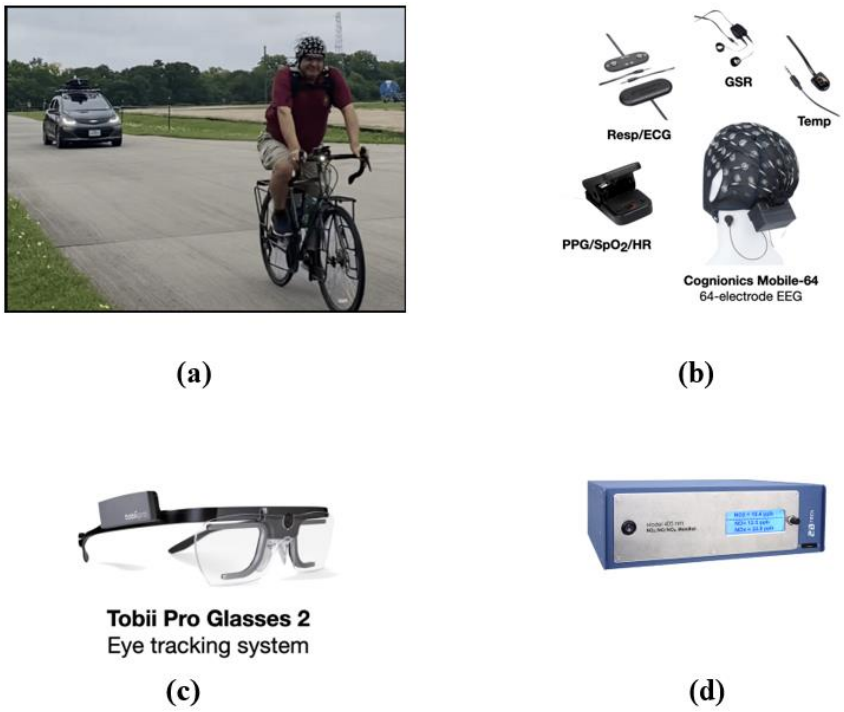

Figure 1 Experimental paradigm and devices used to measure the biometrics of the participant and the ambient NO_2. **(a)** The participant riding a bicycle equipped with a biometric suite to measure biometric variables of the participant and an electric car that follows behind consisting of a NO_2 sensor to measure ambient NO_2. **(b)** Image of the device used to measure cognitive responses using the Cognionics EEG headset and devices to measure some physiological responses such as respiration rate, GSR, ECG, skin temperature, SpO_2, heart rate. **(c)** Image of the Tobii Pro Glasses 2 device used for pupillometric measurements such as distance between the pupils, pupil diameter of the left eye, and pupil diameter of the right eye. **(d)** Device from 2B technologies that was located in the trunk of the car to measure ambient NO_2. Source: Adapted from [16].

Using the measured variables, a total of 329 biometric variables have been considered, among which 320 variables are from the EEG headset device consisting of 64 electrodes (or channels) made by Cognionics (https://www.cgxsystems.com/mobile-128, accessed February 20, 2024) with a sampling rate of 500 Hz. Six physiological responses: skin surface temperature, ECG, SpO_2, GSR, respiration rate, and heart rate were measured using an AIM Generation 2 instrument from Cognionics (https://www.cgxsystems.com/auxiliary-input-module-gen2, accessed 20 February 2024) with a sampling rate of 500 Hz; the image of the devices used is shown in Figure 1b. The rest of the three measured biometric variables are pupillometric measurements: pupil diameter of the right eye, pupil diameter of the left eye, and distance between the pupil which were measured using the Tobii Pro Glasses 2 (https://www.tobii.com/products/discontinued/tobii-pro-glasses-2, accessed 20 February 2024) at a sampling rate of 100 Hz. The image of the Tobii pro Glasses 2 is shown in Figure 1c. The measurement of ambient NO_2 was done using the "Model 405 nm NO_2/NO/NO_x Monitor" from 2B technologies (https://2btech.io/items/other-monitors/model-405-

nm-no2-no-nox-monitor/, accessed 20 Feb 2024). The image of the NO_2 sensor from 2B technologies is shown in Figure 1d. The sampling rate of the instrument is 0.2 Hz or 1 measurement every 5 seconds.

Changes in autonomic physiological responses of the participant cycling outdoors were captured using the biometric suite consisting of several sensors and ambient NO_2 is measured simultaneously. Natural variability and fluctuation in NO_2 levels were observed and no artificial sources were used. An electric car was used so that the sensor measurements taken of NO_2 in the ambient air were not influenced by any car emissions, as there were none. A brief description of the biometric variables is given below.

- Electroencephalography (EEG): EEGs measure the electrical activity on the surface of the brain as a result of the simultaneous activity of groups of neurons. It measures the potential difference (or voltage) between an electrode and a reference electrode. The reference electrode used in this work is a virtual reference that averages the potential of all electrodes. The data received from the device are a time series of voltage. The signal noise that can be induced in the voltages observed on the electrodes due to movements of the head, tongue, jaws, neck, and eyes, blinking, and swallowing were not removed. Data obtained from each electrode as a voltage time series can be transformed from the time domain to the frequency domain, which can be done using the Welch method [20] and was implemented using scipy [21]. The transformation from time domain to frequency domain gives a power spectrum graph with the power spectral density on the Y-axis in units of (V^2/Hz) and the frequency on the X-axis in units of Hz. This frequency is divided into bands: delta (1-3 Hz), theta (4-7 Hz), alpha (8-12 Hz), beta (13-25 Hz), and gamma (25-70 Hz), each corresponding to a state of the brain. Transforming the data as a time series of each 64 electrodes from time domain to frequency domain and dividing the frequency into 5 parts each gives a total of 320 variables from the EEG headset. The codes to retrieve the data and transform from time domain to frequency domain are uploaded in Github and the link is in the supplementary materials.
- Electrocardiography (ECG): An ECG was used to measure the electrical activity of the heart and was measured in units of microvolts. These electrical impulses create contractions in various parts of the heart that maintain blood flow in the human body, and studying these impulses helps to understand the pace and rhythm of the heartbeat, as well as the strength and timing of the measured impulses [22]. The sensor was placed on the upper part of the chest.
- Galvanic Skin Response (GSR): The sweat glands in our body secrete sweat as an involuntary response that can be triggered by factors such as physical exercise, ambient temperature, and in response to stress or emotional stimuli. These sweat glands make the skin more conductive and GSR (or skin conductance) measures the electrical conductivity of the skin [23]. The sensor was placed on the upper back of the participant and measured skin conductivity in units of µSiemens.
- Oxygen saturation (SpO_2): Measures oxygenated hemoglobin compared to deoxygenated hemoglobin [24] as a percentage of saturation. The sensor was placed behind the left ear of the participant.
- Respiration rate: Measured using the GSR device. The respiration rate was measured as the breathing rate per minute.
- Skin temperature: Since a rectal probe core body temperature is extremely uncomfortable to

measure, the temperature of the skin surface was measured where the sensor was placed on the right temple of the participant, measured in units of °C.

- Heart rate: Measured as the number of heart beats per minute using the same device used to measure SpO_2.
- Average pupil diameter: The average pupil diameter of each eyes was calculated using the measured pupil diameter of the left eye and the pupil diameter of the right eye. The units used are millimeters (mm).
- Distance between pupils: Measures the 3-dimensional distance between pupil centers in units of millimeters (mm).
- Difference between pupil diameter: Indicates the difference in the pupil diameters of the two eyes calculated using the pupil diameter of the left eye and the pupil diameter of the right eye. The units used are millimeters (mm).

As the measurement rate of the devices used to measure the biometric variables and ambient NO_2 are different, the entire data set was down sampled to 0.2 Hz or 1 data point every 5 second (the NO_2 observation data rate).

2.2 Data Collection

The biometric variable data collection process was carried out on a single participant due to COVID-19 constraints. To mitigate the issue of the number of participants used, data collection took place on three different days in 2021: May 26, June 9 and June 10 with two trials on each day. The location of data collection was in Breckenridge Park in Richardson, TX. Because of the movement of the sensors, the measurements can sometimes give an erroneous reading and sometimes no reading at all. Moreover, readings for NO_2 from the sensor had additional quality control checks, as the manufacturer outlined, to be considered good quality data.

These criteria included (i) the flow rate of the sample gas entering the sensor to be between 1400 and 1600 cc/min, (ii) the ozone flow rate to be between 60 and 80 cc/min, (iii) the cell photodiode voltage (PDV) to be at least 0.6 volt, and (iv) the PDV ozone generator to have a voltage of at least 0.1. Data were cleaned using the Pandas library [25, 26] which reduced the number of data points. The process of data collection, down sampling and data cleaning gave a total of 582 data points with: 136, 23, 126, 120, 132, 45 data points for Trail 1, Trial 2, Trial 3, Trial 4, Trial 5 and Trial 6 respectively, essentially a dataset with 582 rows and 330 columns of which 329 columns are the biometric input variables, and the last column is the target variable NO_2 we would like to estimate and the 582 rows are the measurements made for 582 discrete time steps.

The location of the bike ride was measured using a GPS sensor on the bike. Figure 2 shows the location where the data collection took place for Trials 2, 3, 4, 5, and 6. As the GPS sensor did not work during Trial 1, no exact location was monitored but the route used was the same as that of Trial 3 and Trial 5. Each of the circular dots in the subfigure shows the spot where all biometric variables and ambient NO_2 were measured simultaneously with the corresponding ambient NO_2 concentration in the color map. The arrows in the subfigures indicate the initial direction of the ride. Variation in NO_2 levels can be observed in all five trials, which could be due to nearby traffic emissions.

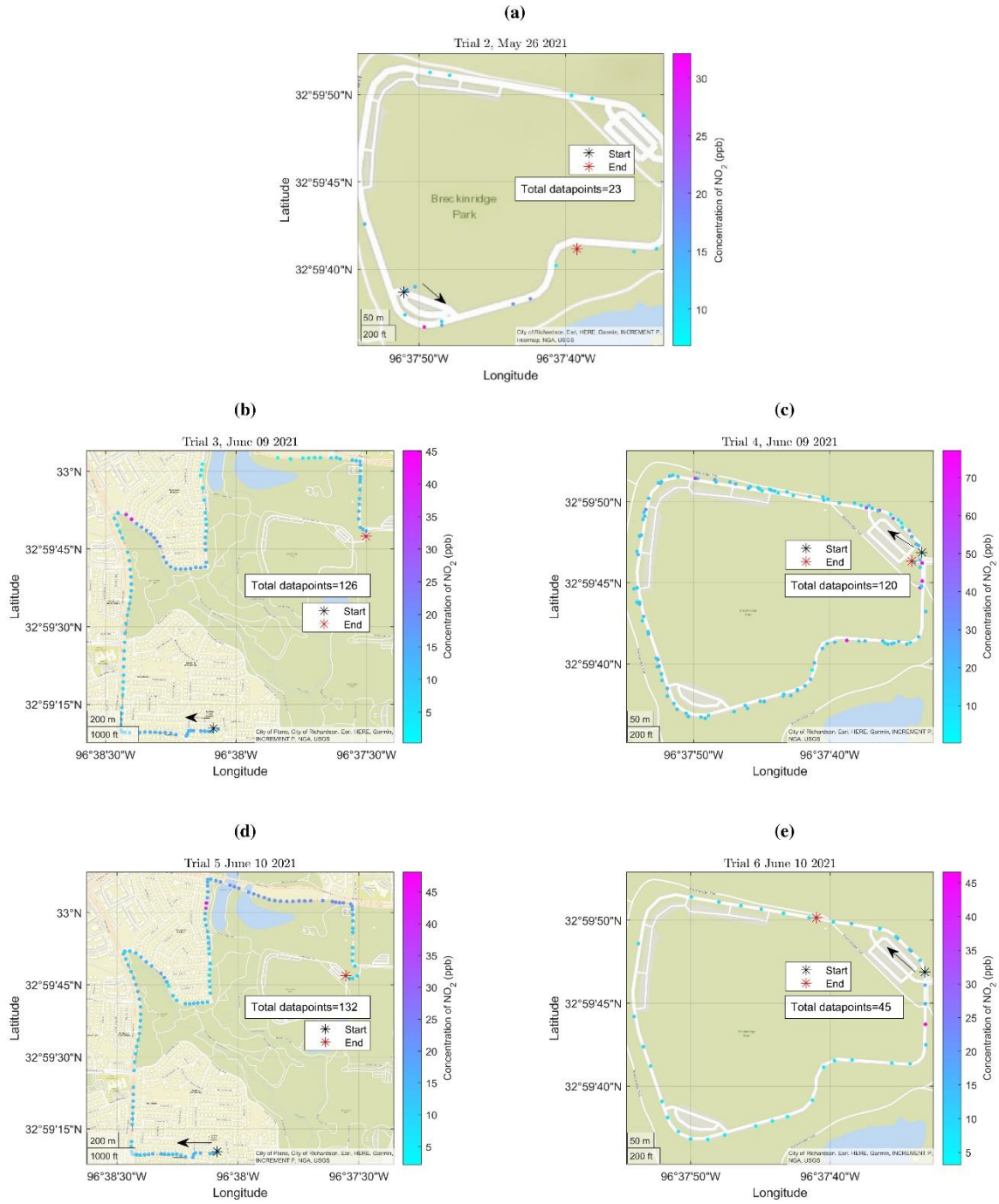

Figure 2 Location of data collection for 5 of the 6 trials. Each of subfigure **(a)**, **(b)**, **(c)**, **(d)** and **(e)** shows the location and the places where the measurement of biometric variables and ambient NO_2 was performed simultaneously with the corresponding value of NO_2 on the color map. The arrows indicate the initial direction of the bicycle ride.

Since all data were quality controlled and downsampled at one data point every 5 seconds, we can sometimes see a discontinuous path rather than a continuous path in each of the trials, which is prominent in trial 2 in Figure 2a, while an almost continuous path can be seen in Figures 2b-2d.

2.3 Data Analysis and Machine Learning Model Development

The estimation of inhaled NO_2 using biometric variables was performed using a Random Forest [27] algorithm for nonlinear, nonparametric, multidimensional regression using the Ensemble Random Forest Regressor package from scikit-learn [28] using default parameters. The biometric variables were used as input features for the machine learning model to estimate the predictor variable, which in this case is the ambient air concentration of inhaled NO_2. The estimated values are then compared with the true values measured by the NO_2 sensor placed in the electric car. 80% of the data was used to train the machine learning model, the remaining 20% was used as an independent test set. To quantify the precision of the prediction, the determination coefficient (R^2) and the root mean square error (RMSE) are calculated between the actual and estimated values of NO_2. The scatter plot and the Quantile-Quantile plot between the actual and estimated values of NO_2 are also plotted for a qualitative test of prediction precision. A time series graph of the actual values of NO_2 overlaid with the estimated values of NO_2 is also plotted for a qualitative analysis of the prediction.

To identify the biometric variables that were the most important or contributed the most to the estimation of inhaled NO_2, SHAP values (SHapley additive explanations) [29, 30] were used to classify the effectiveness of the biometric variables in descending order. Combining the top 9 predictor variables identified using SHAP values and the target variable NO_2, a 10 × 10 Pearson correlation coefficient matrix is calculated to identify the linear relationship between the variables. A 10 × 10 matrix of mutual information of the same variables is also calculated to capture the nonlinear relationship between the variables using a package from scikit-learn [28]. These values of mutual information are zero for variables that are independent of each other, and the number keeps on increasing if the relationship is stronger but typically is below 5.

All experimental protocols were approved by the Institutional Review Board of the University of Texas at Dallas and informed consent was received from the participant.

3. Results

The process of estimating the inhaled concentration of NO_2 from the biometric variables using machine learning was carried out in three different ways: first, we used all 329 biometric variables that were measured or calculated. The machine learning model in which 329 biometric variables are used to estimate inhaled NO_2 is quite complex as the number of input parameters is large, especially considering the small number of time points collected due to the constraints placed during COVID. Therefore, in the second machine learning model, the number of biometric variables used to estimate NO_2 was reduced and only the nine physiological responses were used to estimate inhaled NO_2. In the third and final model, the number of biometric variables was reduced to just six, chosen based on the importance ranking provided by the SHAP value plot to make the machine learning model even simpler, following Occam's razor.

3.1 Using 329 Features

Using 329 features as input features to estimate inhaled NO_2 using the Random Forest algorithm from scikit-learn, the results in the training data are very high, which is to be expected since this part of the data set is used by the algorithm for learning. The coefficient of determination (R^2) and

RMSE between the actual and estimated values of NO_2 were 0.92 and 2.90 ppb, respectively. Figure 3a shows a bar graph of the RMSE values in the training set in blue and those in the testing set in orange. However, the results in the testing set were not high as R^2 and RMSE between the actual and the estimated values of inhaled NO_2 were found to be 0.16 and 9.62 ppb, respectively. It is to be noted that since the model is complex with a large number of features, these numbers, however, do change depending on how the data is shuffled.

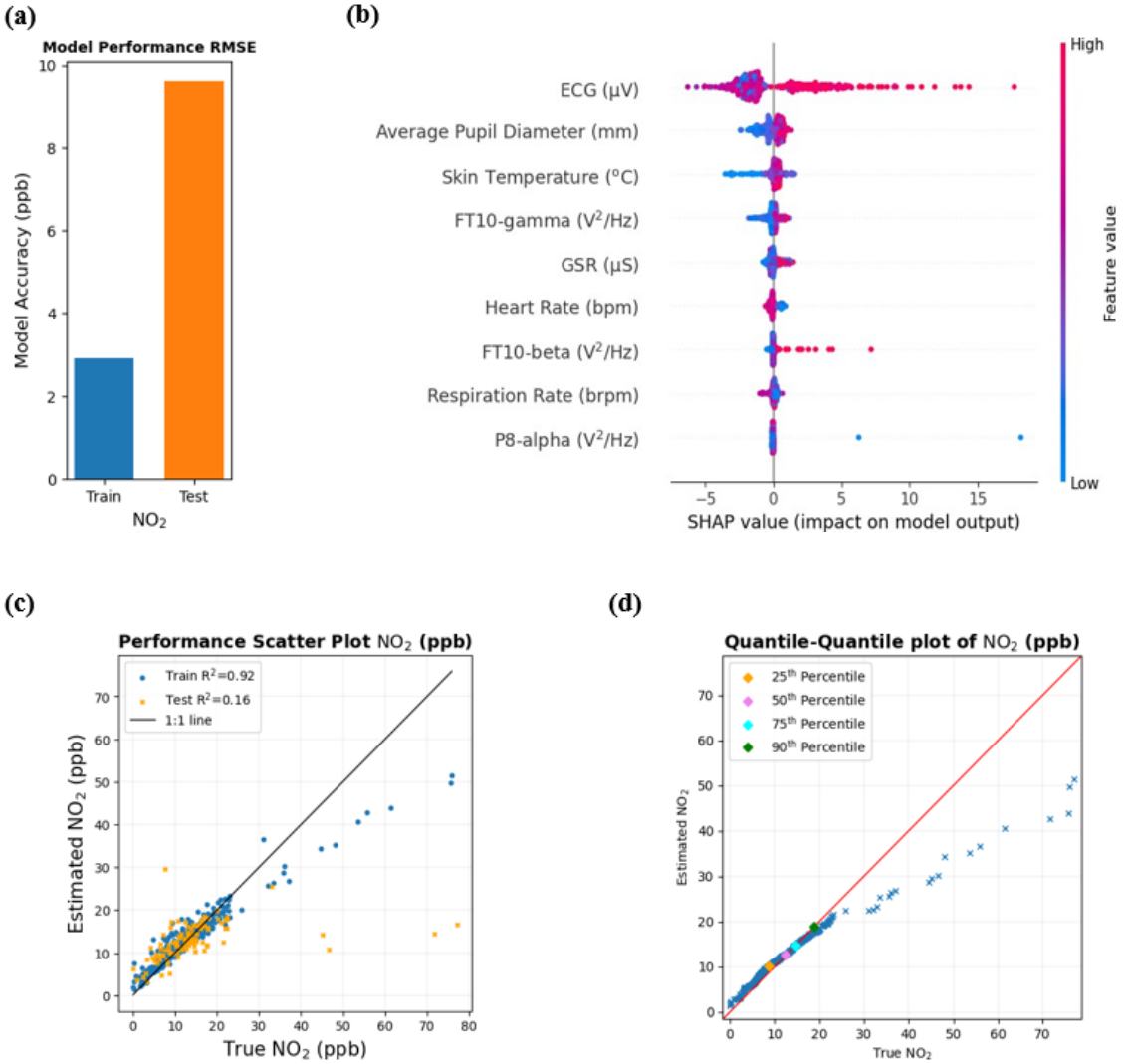

Figure 3 Model performance and top 9 feature importance plot for estimating inhaled NO_2 using 329 biometric features. **(a)** RMSE between the actual and estimated values of NO_2 in the training and testing set. **(b)** Top 9 features in estimating NO_2 as identified by SHAP values plotted in a beeswarm plot. **(c)** Scatter plot between the actual and estimated values in the training and the test set. **(d)** Quantile-Quantile plot between the actual and estimated values of NO_2 with the overlaid percentiles.

Figure 3c shows a scatter plot between the true values of NO_2 and the estimated values of NO_2 with the data points used in training represented by blue circular dots and the data points in the test set represented by the orange "x" sign with a black 1:1 line overlaid. Data points with an exact prediction will be in the black 1:1 line. The diagram shows that for smaller values of NO_2 where there is an abundance of data, most of the data points lie close to the 1:1 line, while for higher values of

NO_2 where there is scarcity of data points both in the training set and the testing set, the data points deviate from the 1:1 line. Figure 3d shows a Quantile-Quantile graph with true values of NO_2 on the X-axis and estimated values of NO_2 in the Y-axis overlaid with a red 1:1 line. Data points with an exact prediction will be on the red 1:1 line. The Quantile-Quantile plot in Figure 3d also shows the same results. It can be seen that the quantiles are very close to the red 1:1 line for over 90% of the data. On the other hand, for higher values of NO_2 the quantiles deviate to a large extent from the 1:1 red line. Figure 3b shows the SHAP values of the top 9 features in a beeswarm plot.

As the SHAP value for the ninth feature is nearly zero, all biometric variables below have even smaller SHAP values and therefore have almost no contribution to the estimation of the predictor variable NO_2. Each of the dots represents the SHAP value of the corresponding data of the feature, so each biometric variable has 582 circular dots in the diagram. A color map is shown on the right side with high feature values in red and low feature values in blue. X-axis of the plot shows the SHAP value in ppb of the corresponding value of the feature. The biometric variables are arranged in descending order according to the absolute value of the SHAP values of the feature. The magnitude of SHAP values for ECG, average pupil diameter, and skin temperature is higher compared to rest of the features, so the ranking of these three features tends to remain consistent. However, as the model is complex and the SHAP values for some of these features are close to each other, the ordering changes a little depending on how the data are shuffled. The figure shows only one instant when the algorithm was run. One of the most important features for estimating inhaled NO_2 was found to be the ECG, which is expected, as environmental NO_2 has been associated with cardiovascular issues [11, 31]. The figure also shows some of the other physiological responses that were crucial to estimate inhaled NO_2 which include skin temperature and GSR. These two biometric variables were also some of the important variables that were useful in predicting PM_1, $PM_{2.5}$ [16] and CO_2 [17] using biometric variables. Heart rate and respiration rate were also some of the other important features. Since the participant is cycling, which involves a lot of physical work, there will be changes in body temperature, sweating, heart rate, and respiration rate. However, inhalation of NO_2 and other components of air quality is known to affect the respiratory system, create cardiovascular problems, inflammation of the airways and possibly cause these autonomous physiological responses.

Other biometric variables in the list of the top 9 features include some EEG electrodes. As electrodes are named according to the 10-10 nomenclature system [32], the location of the electrode can be determined by its name. The FT10 electrode is located between the frontal and temporal lobes on the right side of the brain. The frontal lobe is involved in tasks such as making decisions and movement, while the temporal lobe is involved in tasks such as speech, musical rhythm, short-term memory, and smell [33]. Similarly, the P8 electrode located on the left side of the brain in the parietal lobe seems to have a very small magnitude of SHAP value, therefore all other features below it have lower SHAP values, therefore having less of a contribution to predict inhaled NO_2.

The 10 × 10 Pearson correlation coefficient matrix in Figure 4a shows that most of the variables are not linearly related with each other and do not have linear relation with the target variable as well. The 10 × 10 mutual information matrix of the same variables in Figure 4b, shows the non-linear relation between the variables.

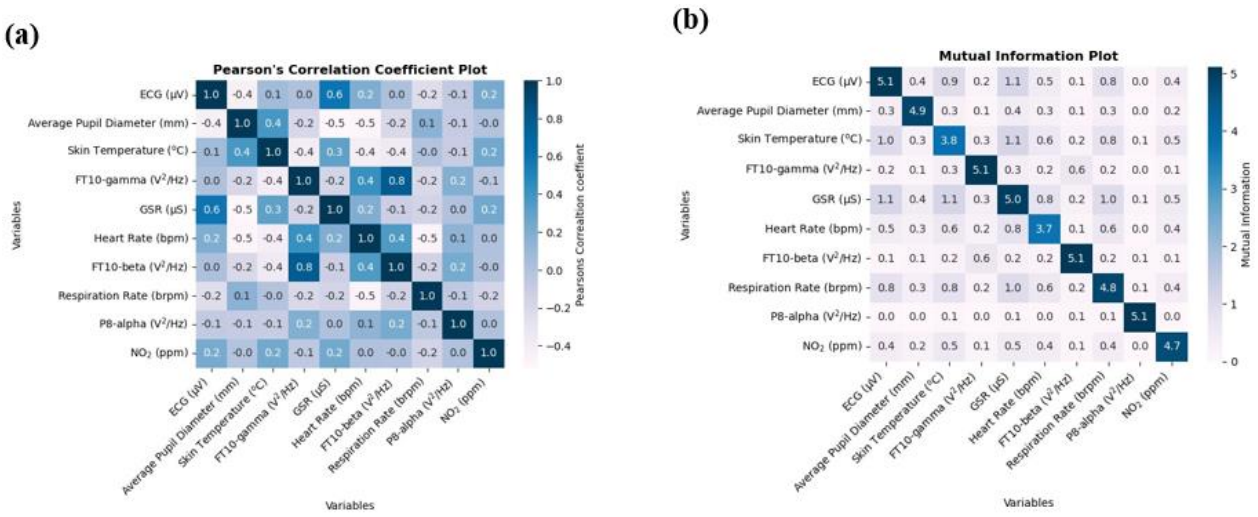

Figure 4 10 × 10 Pearson's correlation and mutual information matrix **(a)** Pearson's correlation matrix for 10 variables to identify linear correlation. **(b)** Mutual information matrix for 10 variables to identify linear and non-linear relation.

Figure 4b, shows that skin temperature, GSR, heart rate, and respiration rate had higher mutual information with NO_2 which is to be expected as the SHAP values for these features were also high. There was nonlinear relationship between the biometric variables as well, particularly, ECG with skin temperature, GSR, respiration rate; skin temperature with GSR, respiration rate; GSR with heart rate, respiration rate; heart rate with respiration rate depicting that the biometric variables are mutually related with each other.

3.2 Using a Reduced Number of Features

We now simplify the machine learning model by reducing the number of features to estimate inhaled NO_2. First, we consider all of the 9 physiological responses and then select only 6 of the top features based on SHAP values.

The results obtained from reducing the number of features to estimate inhaled NO_2 from 329 to 9 appear to be similar. The coefficient of determination (R^2) between the true and estimated values of NO_2 in the train and the test set was 0.92 and 0.33 respectively. As shown in the bar graph in Figure 5a, the RMSE between the true and estimated values of NO_2 in the training set was 2.85 ppb and 5.90 ppb, respectively. This is similar to the case where all 329 biometric variables were used to estimate inhaled NO_2. ECG, respiration rate, and skin temperature are some of the physiological responses that were most effective in estimating inhaled NO_2 as shown by the ranking of these variables in the beeswarm plot in Figure 5b, which is similar compared to Figure 3b. The effectiveness of variables such as the average pupil diameter, SpO_2 and pupil distance seems to be small in estimating inhaled NO_2 as their SHAP values are small compared to other features as shown in Figure 5b. While these results can change a little based on how the data is shuffled, now that the model is simpler, these results will be more or less consistent every time the algorithm is run.

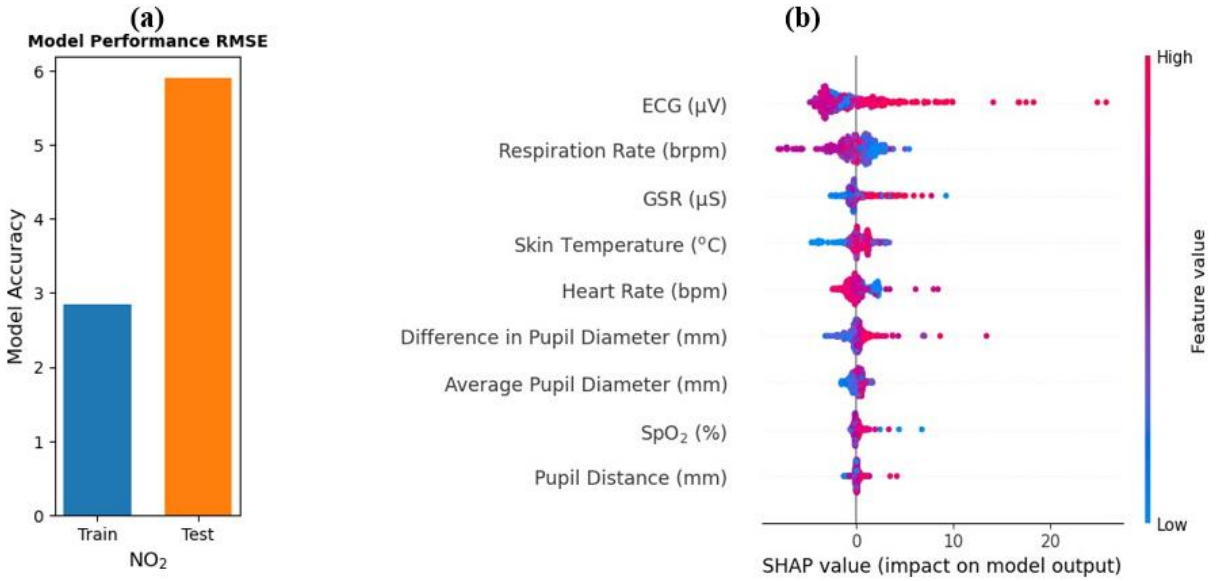

Figure 5 Performance and top 9 feature importance plot for estimating NO_2 using 9 physiological responses. **(a)** RMSE between the actual and estimated values of NO_2 in the training and testing set. **(b)** Top 9 features for the estimation of NO_2 identified using SHAP values and plotted in a beeswarm plot.

Since features such as the average diameter of the pupils, SpO_2 and pupil distance had little effect on the estimation of inhaled NO_2, we now test the accuracy of the prediction by removing these 3 features among the 9 features. Six of the features now used to estimate inhaled NO_2 are: ECG, respiration rate, skin temperature, difference in pupil diameter, GSR, and heart rate. The determination coefficient (R^2) and the RMSE between the true values and estimated values of NO_2 are 0.35 and 5.41 ppb in the test set. The results are again similar when 9 physiological responses were taken into account for the estimation of inhaled NO_2.

Table 1 shows the summary of the results obtained. The determination coefficient (R^2) and the RMSE between the true values of NO_2 and the estimated values of NO_2 in the training and testing set are shown with the corresponding number of biometric variables used to make the prediction.

Table 1 Summary of quantifying the accuracy of the estimation of inhaled NO_2 with the corresponding number of features used.

Number of biometric variables	Trian R^2	Test R^2	Train RMSE	Test RMSE
329	0.92	0.16	2.90 ppb	9.62 ppb
9	0.92	0.33	2.85 ppb	5.90 ppb
6	0.92	0.35	3.05 ppb	5.41 ppb

From Table 1, we can see that the results are similar and sometimes even improve (when the algorithm is rerun) even considering only a subset of variables. This in fact does align with the Occam's razor principle, and we can select just a part of these variables to make the model simpler, which usually generalizes well.

A scatter plot and Quantile-Quantile plot by considering 9 physiological responses and by considering 6 physiological responses to estimate inhaled NO_2 are shown in Figure 6.

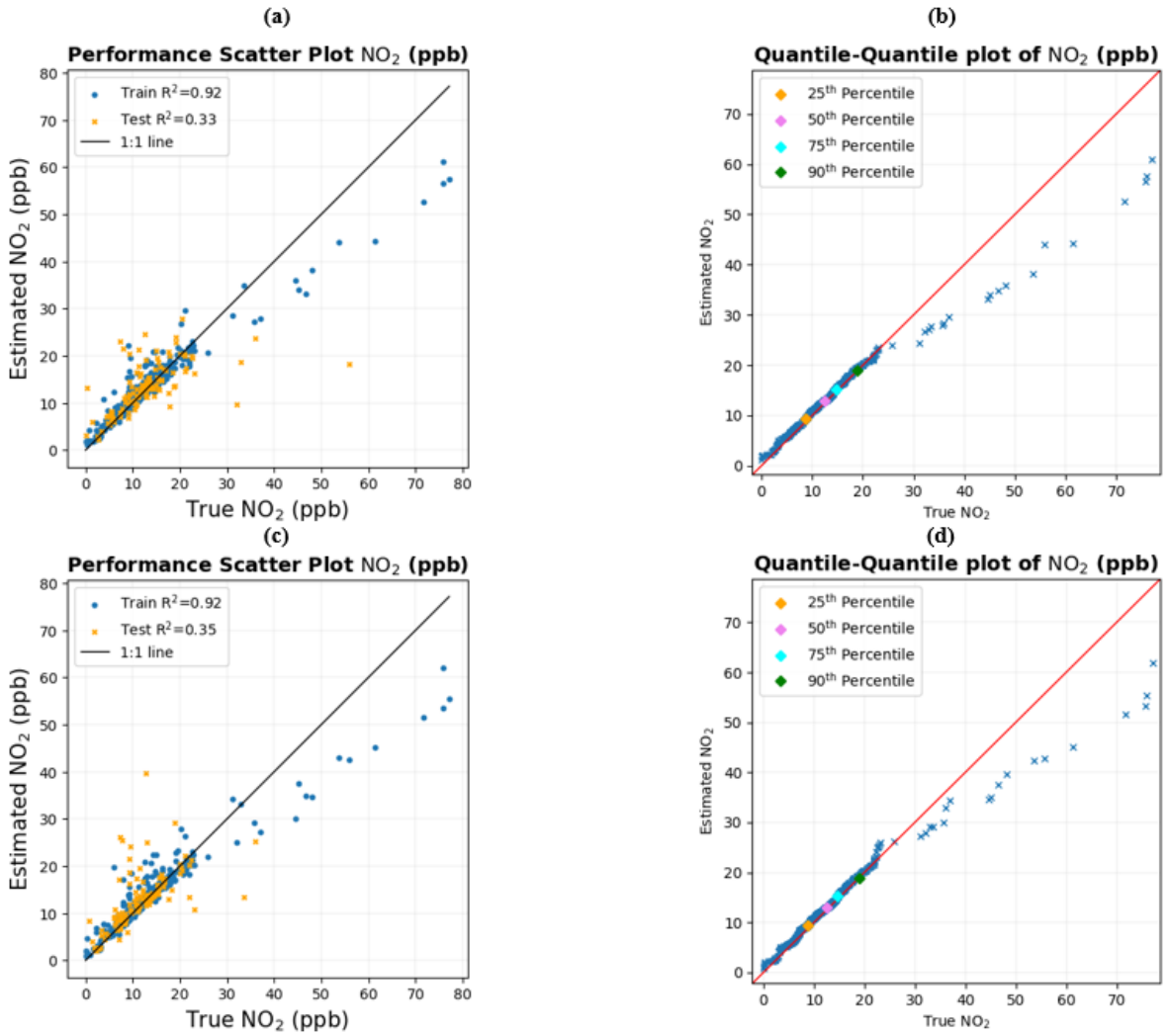

Figure 6 Scatter and quantile-quantile plot between true and estimated values of NO_2. **(a)** Scatter plot between the actual and estimated values in the training and the test set considering 9 physiological responses. **(b)** Quantile-Quantile plot between the actual and estimated values considering 9 physiological responses with the percentiles of the distribution overlaid. **(c)** Scatter plot between the actual and estimated values in the training and the test set considering 6 physiological responses. **(d)** Quantile-Quantile plot between the actual and estimated values considering 6 physiological responses with the percentiles of the distribution overlaid.

Figure 7 shows a time series of the actual values of NO_2 overlayed with the estimated values of NO_2 when only six physiological responses were used for the prediction of inhaled NO_2. True values of NO_2 are shown on a continuous orange line, while the estimated values of NO_2 are shown on a dotted blue line. The shaded background represents different trials for each of the six trials. Lines are stopped whenever trials differ. It can be seen that for most time series, especially when the NO_2 values are small, where there is a large amount of data, the true values of NO_2 are close to the estimated values.

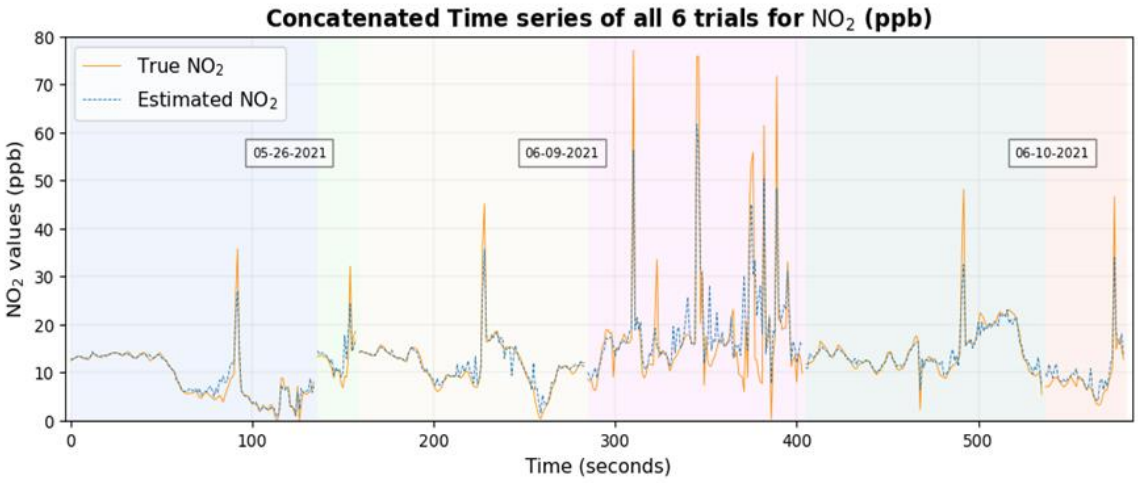

Figure 7 Time series plot of the true NO_2 values with the estimated values of NO_2 overlaid for all the 6 trials of data collected on 3 separate days with 2 trials on each day when 6 biometric variables were used for the prediction of inhaled NO_2.

4. Discussion

Measurements of autonomic physiological responses made on small temporal and spatial scales coupled with the use of machine learning models to predict particulate matter, in particular, PM_1 and $PM_{2.5}$ and even gases such as CO_2 were found to be an effective methodology [16-18]. The basis of this study is to investigate whether autonomous physiological and cognitive changes induced by air pollution components such as NO_2 can be used to estimate and understand the effects of inhaled NO_2. The results when 9 and 6 biometric variables were used to estimate inhaled NO_2 appear to be moderate as indicated by the coefficient of determination (R^2) and the RMSE values between the estimated and true values of NO_2 as shown in Table 1. NO_2 as an air pollution component is known to have many health-related effects [11-15], the use of biometric variables to estimate the gas itself seems to be an approach in the right direction.

However, the result of the estimation is not as accurate as those for CO_2 and particulates, which could be attributed to two possible reasons:

- Machine learning models require a large number of training data to learn the parameters that minimize error in the best possible way. The scatter diagram in Figure 3c, Figure 6a and Figure 6c shows that most of the data points lie close to the black line 1:1 where the NO_2 values are small and there is an abundance of data points, while the number of data points in the training set for NO_2 values above 30 ppb is less than 15 and the number of data points in the test set is less than 5. As the value of NO_2 increases, there is a scarcity of data points for the machine learning model to learn from the training data and then test on the test data points. Thus, the data points begin to deviate from the 1:1 line. A similar result is also seen in the Quantile-Quantile graph in Figure 3d, Figure 6b and Figure 6d. The Quantile-Quantile plot shows that more than 90% of the data points are below 20 ppb. Below this margin, the Quantile-Quantile plot is very close to the 1:1 red line while points deviate from the red 1:1 line as the number of data points decreases for higher values of NO_2.
- The concentration of PM_1 particles in ambient air for the tests on 3 separate days ranged between 0.708 to 7.655 $\mu g/m^3$ whereas the concentration of NO_2 ranged from 0.1 to 77.1 ppb.

These PM_1 particles, with their small size, can easily mix in the air and penetrate deep into the lungs and bloodstream. Therefore, autonomic responses were likely dominated by these PM_1 particles rather than NO_2, which made the estimate not as high as that obtained for PM_1 where $R^2 = 0.91$ [16].

Since the prediction of NO_2 was observed to be accurate for smaller values of NO_2 as qualitatively indicated by the scatter diagram in Figure 3c, Figure 6a and Figure 6c and the Quantile-Quantile plots in Figure 3d, Figure 6b and Figure 6d, the second plausible reason is probably the reason that the precision is not high in estimating NO_2 for the entire data set. This hypothesis is also supported by the time series plot in Figure 7 where the true values of NO_2 are close to the estimated values of NO_2 for lower values of NO_2. These results suggest an abundance of data points which can be achieved in multiple ways, such as (a) the use of an instrument that measures ambient NO_2 with a high frequency of data collection. (b) The use of multiple participants and collecting data over a long period of time can possibly improve the results of the prediction.

The use of Occam's razor principle to simplify the model, which generally generalizes well, seems to be well aligned in this case, as indicated by the R^2 and RMSE values between the true and estimated values of NO_2 as shown in Table 1. These metrics are similar when the number of features was reduced. The scatter diagram in Figure 3c, Figure 6a and Figure 6c and the Quantile-Quantile plots in Figure 3d, Figure 6b and Figure 6d also shows the overall structure of these plots are similar. Therefore, rather than using a large set of biometric variables, making use of a subset of variables seems to be an efficient way to estimate inhaled NO_2.

Two of the limitations in this study include one being the number of participants and the other being the noise in the EEG data. This study included data measurements on a single participant. To generalize the result, data collection should be carried out on a larger number of participants with an abundance of data over a range of NO_2 values. Also, EEG signal measurement was done when the participant was doing physical work resulting in frequent blinking, jaw clenching, tongue movement, etc. which distorts the EEG signal as mentioned before. Removing these artifacts, which most of the time can be a combination of artifacts, is a challenging task on its own.

Extension of this work involves the collection of large number of data as required by machine learning models which can be achieved by procedures mentioned before. This may possibly significantly increase the prediction. Better machine learning models, which can train and test on a small number of data points, could also make the prediction better for the current dataset and is also another approach that could possibly provide better results. Furthermore, since the variation of NO_2 was dependent on ambient air, confounding variables are expected, which can be studied using casual analysis. To better understand the direct effects of NO_2 inhalation, participants could also be placed indoors with artificial variation of NO_2. The methodology could also be used to test the effects and estimation of other pollutants such as carbon monoxide.

5. Conclusion

The methodology of making use of machine learning for regression to estimate inhaled NO_2 by using the measured autonomic response on a small temporal and spatial scale in microenvironments seems to be effective for small values of NO_2 under 20 ppb where 90% of the data set was most abundant but the precision was only moderate for the entire data set as the coefficient of determination and RMSE between the actual and estimated values were found to be

0.35 and 5.41 ppb, respectively, in the test set. Large numbers of data collected from multiple participants over a range of target variables can be used so that machine learning models can be better trained and then to be tested on an independent test set. Machine learning models that can work on a limited set of data could possibly improve the results as well. Furthermore, instead of making use of a large number of biometric variables, which in some cases tends to overfit the data, a subset of variables can be used to make the model simpler, which could generalize well.

Abbreviations

The abbreviations used in the study are as follows:

ppb	Parts per billion
EEG	Electroencephalography
GSR	Galvanic Skin Response
ECG	Electrocardiography
SpO_2	Blood Oxygen Saturation
PM	Particulate Matter
RMSE	Root Mean Square Error

Acknowledgments

The authors highly acknowledge the support that was received from the University of Texas at Dallas Office of Sponsored Programs, Dean of Natural Science and Mathematics of the University, and Chair of the Physics Department.

Author Contributions

Methodology D.J.L., S.T. and T.L.; software S.T., S.R.; formal analysis D.J.L., S.R.; data curation S.T., D.J.L., L.W., B.A.F., T.L., M.L., J.S., A.A. and J.W.; writing-original draft preparation S.R. and D.J.L; writing review and editing, S.R., D.J.L., P.M.H.D.; visualization S.R.; supervision D.J.L.

Funding

The following grants were helpful in this work: The US Army (Dense Urban Environment Dosimetry for Actionable Information and Recording Exposure, US Army Medical Research Acquisition Activity, BAA CDMRP Grant Log #BA170483. The Texas National Security Network Excellence Fund Award for Environmental Sensing Security Sentinels. EPA 16th annual P3 Awards Grant Number 83996501, entitled Machine Learning-Calibrated Low-Cost Sensing. SOFTWERX award for Machine Learning for Robotic Teams. TRECIS CC* Cyberteam (NSF #2019135); NSF OAC-2115094 Award; and EPA P3 grant number 84057001-0.

Competing Interests

The authors have declared no competing interests.

Additional Materials

The code and data are publicly available and are available on GitHub: https://github.com/mi3nts/Estimate-inhaled-NO2. The entire data set is also available in Zenodo: https://zenodo.org/records/10345982 (accessed on December 11,2023).

References

1. Shaddick G, Thomas ML, Mudu P, Ruggeri G, Gumy S. Half the world's population are exposed to increasing air pollution. NPJ Clim Atmos Sci. 2020; 3: 23.
2. World Health Organization. Air pollution [Internet]. Geneva, Switzerland: World Health Organization; 2023. Available from: https://www.who.int/health-topics/air-pollution#tab=tab_1.
3. Anderson JO, Thundiyil JG, Stolbach A. Clearing the air: A review of the effects of particulate matter air pollution on human health. J Med Toxicol. 2012; 8: 166-175.
4. Sacks JD, Stanek LW, Luben TJ, Johns DO, Buckley BJ, Brown JS, et al. Particulate matter-induced health effects: Who is susceptible? Environ Health Perspect. 2011; 119: 446-454.
5. Jacobson TA, Kler JS, Hernke MT, Braun RK, Meyer KC, Funk WE. Direct human health risks of increased atmospheric carbon dioxide. Nat Sustain. 2019; 2: 691-701.
6. Hanley ME, Patel PH. Carbon monoxide toxicity. Treasure Island, FL: StatPearls Publishing; 2023.
7. Ritz B, Hoffmann B, Peters A. The effects of fine dust, ozone, and nitrogen dioxide on health. Dtsch Arztebl Int. 2019; 116: 881-886.
8. Orellano P, Reynoso J, Quaranta N. Short-term exposure to Sulphur dioxide (SO_2) and all-cause and respiratory mortality: A systematic review and meta-analysis. Environ Int. 2021; 150: 106434.
9. Raj K, Das AP. Lead pollution: Impact on environment and human health and approach for a sustainable solution. Environ Chem Ecotoxicol. 2023; 5: 79-85.
10. American Lung Association. Nitrogen dioxide [Internet]. Chicago, IL: American Lung Association; 2023. Available from: https://www.lung.org/clean-air/outdoors/what-makes-air-unhealthy/nitrogen-dioxide.
11. Samoli E, Aga E, Touloumi G, Nisiotis K, Forsberg B, Lefranc A, et al. Short-term effects of nitrogen dioxide on mortality: An analysis within the APHEA project. Eur Respir J. 2006; 27: 1129-1138.
12. Gillespie-Bennett J, Pierse N, Wickens K, Crane J, Howden-Chapman P. The respiratory health effects of nitrogen dioxide in children with asthma. Eur Respir J. 2011; 38: 303-309.
13. Latza U, Gerdes S, Baur X. Effects of nitrogen dioxide on human health: Systematic review of experimental and epidemiological studies conducted between 2002 and 2006. Int J Hyg Environ Health. 2009; 212: 271-287.
14. Cibella F, Cuttitta G, Della Maggiore R, Ruggieri S, Panunzi S, De Gaetano A, et al. Effect of indoor nitrogen dioxide on lung function in urban environment. Environ Res. 2015; 138: 8-16.
15. Brender JD. Human health effects of exposure to nitrate, nitrite, and nitrogen dioxide. In: Just enough nitrogen. Cham: Springer; 2020. pp. 283-294.
16. Talebi S, Lary DJ, Wijeratne LO, Fernando B, Lary T, Lary M, et al. Decoding physical and cognitive impacts of particulate matter concentrations at ultra-fine scales. Sensors. 2022; 22: 4240.

17. Ruwali S, Fernando BA, Talebi S, Wijeratne L, Waczak J, Sooriyaarachchi V, et al. Gauging ambient environmental carbon dioxide concentration solely using biometric observations: A machine learning approach. Med Res Arch. 2024; 12. doi: 10.18103/mra.v12i1.4890.
18. Fernando BA, Talebi S, Wijeratne L, Waczak J, Sooriyaarachchi V, Ruwali S, et al. Data-driven environmental health: Unraveling particulate matter trends with biometric signals. Med Res Arch. 2024; 12. doi: 10.18103/mra.v12i1.4899.
19. Lary DJ, Faruque FS, Malakar N, Moore A, Roscoe B, Adams ZL, et al. Estimating the global abundance of ground level presence of particulate matter ($PM_{2.5}$). Geospat Health. 2014; 8: S611-S630.
20. Welch P. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. IEEE Trans Audio Electroacoust. 1967; 15: 70-73.
21. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nat Methods. 2020; 17: 261-272.
22. John Hopkins Medicine. Electrocardiogram [Internet]. Baltimore, MD: John Hopkins Medicine; 2023. Available from: https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/electrocardiogram.
23. Albert WB, Tullis TST. Chapter 8 - measuring emotion. In: Measuring the user experience. 3rd ed. Burlington, MA: Morgan Kaufmann; 2023. pp. 195-216.
24. Jubran A. Pulse oximetry. Crit Care. 1999; 3: R11.
25. The pandas development team. pandas-dev/pandas: Pandas. Geneva, Switzerland: Zenodo; 2024. doi: 10.5281/zenodo.10957263.
26. McKinney W. Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference (SciPy 2010); 2010 July 3-28; Austin, TX, USA. SciPy.org. doi: 10.25080/Majora-92bf1922-00a.
27. Breiman L. Random forests. Mach Learn. 2001; 45: 5-32.
28. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011; 12: 2825-2830.
29. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 December 4-9; Long Beach, CA, USA. Red Hook, NY: Curran Associates Inc.
30. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020; 2: 56-67.
31. Faustini A, Rapp R, Forastiere F. Nitrogen dioxide and mortality: Review and meta-analysis of long-term studies. Eur Respir J. 2014; 44: 744-753.
32. Acharya JN, Hani AJ, Cheek J, Thirumala P, Tsuchida TN. American clinical neurophysiology society guideline 2: Guidelines for standard electrode position nomenclature. Neurodiagn J. 2016; 56: 245-252.
33. Jawabri KH, Sharma S. Physiology, cerebral cortex functions. Treasure Island, FL: StatPearls Publishing; 2023.