Medical Research Archives

European Society of Medicine

OPEN ACCESS

Published: January 31, 2024

Citation: Ruwali, S., et al., 2024. Gauging Ambient Environmental Carbon Dioxide Concentration Solely Using Biometric Observations: A Machine Learning Approach. Medical Research Archives, [online] 12(1). https://doi.org/10.18103/mra.v12i1.4890

DOI: https://doi.org/10.18103/mra.v12i1.4890

ISSN: 2375-1924

RESEARCH ARTICLE

Gauging Ambient Environmental Carbon Dioxide Concentration Solely Using Biometric Observations: A Machine Learning Approach

Shisir Ruwali, Bharana Ashen Fernando, Shawhin Talebi, Lakitha Wijeratne, John Waczak, Vinu Sooriyaarachchi, Prabuddha Hathurusinghe, 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.

*david.lary@utdallas.edu

ABSTRACT

Respiration is vital for human function. Inhaling specific gases can have specific physiological and cognitive impacts. Using a suite of sensors, we can collect detailed information on a range of both physiological and environmental factors. This study builds on previous research exploring how particulate matter affects physiological and cognitive responses, now expanded to include CO_2. We tracked the biometric variables of a cyclist, analyzing 329 specific variables. Simultaneously, an electric vehicle following the cyclist measured CO_2 and other environmental factors. After data collection, we used machine learning models to decipher the interactions between the human body and its surroundings. We found that biometric data alone could be used to accurately estimate the amount of CO_2 inhaled, achieving a good level of precision (r^2=0.98) when comparing the estimated CO_2 based on biometrics and the actual observed CO_2 levels. In addition, we developed a ranking system to identify the biometric variables that most significantly predict environmental CO_2 inhalation.

Keywords: Machine learning, biometric, particulate matter, cognitive, CO_2.

1. Introduction

During inhalation, air containing oxygen and other components is drawn into our lungs. As a byproduct of our body's metabolism, carbon dioxide is produced and enters the lungs via the bloodstream. This waste gas is then expelled from the lungs during exhalation[1]. Recent data from the World Health Organization reveal that 99% of the global population is breathing air that exceeds WHO quality guidelines. This widespread exposure to air pollution at home and in the environment contributes to approximately 7 million premature deaths per year[2].

Since respiration is an important part of life, it is natural that the quality of air we breathe has many effects on the human body[3]. Exposure to gases such as sulfur dioxide, ozone, nitrogen oxides, and carbon monoxide negatively affects the respiratory, cardiovascular and other systems of the human body, while particles such as lead and mercury can negatively affect the nervous, urinary, and cellular mechanisms, among others[4,5,6,7,8,9,10].

This study continues the work done previously to understand the effects of air pollution, particularly particulate matter in the human body[11], expanding it to examine the autonomic response in microenvironments at small spatial and temporal scales by now considering inhaled CO_2.

Previous studies have shown that even short-term exposure to CO_2 can lead to various physiological and cognitive effects with a concentration of CO_2 ranging from 500 ppm to 3,000 ppm[12,13,14]. Short-term exposure, but with a high concentration of CO_2 in the range of 7%-14% has been found to increase arterial pressure, heart rate, and gas volume intake with other effects such as abnormal cardiac rhythm, sweating, headache, and auditory and visual problems[15].

The purpose of this work is to simultaneously sense both environmental and biometric data and then examine the utility of machine learning to build empirical models of the observed interactions. As the functional forms of the relationships between the observed variables are not always known, machine learning techniques can be used to learn by example these multi-variate and often nonlinear relationships. The biometric variables observed simultaneously include nine physiological responses, namely heart rate (HR), galvanic skin response (GSR), respiration rate (RR), skin temperature, blood oxygen saturation (SpO_2), electrocardiography (ECG), average diameter of the pupil of both eyes, absolute value of the difference in pupil diameter, the distance between pupils, and measurement of electrical activity at 64 locations throughout the head using electroencephalography (EEG). Among these variables, the absolute value of the difference in pupil diameter and average pupil diameter are calculated variables using the measured values of diameter of both eyes of the participant, while the EEG and seven other physiological variables are directly measured.

Since inhaled CO_2 affects our body in multiple ways, we investigated whether these biometric variables can be used to estimate the concentration of inhaled CO_2. We were able to use machine learning regression to estimate inhaled CO_2 from just a subset of the observed biometric data. After training the model, which used 80% of the complete data set, we could test how the model performs using the remaining data as an independent

validation data set. It turns out that, for this purpose, some biometric variables are much more useful than others, so we can use only a subset of biometric variables to accurately estimate the inhaled environmental CO_2.

2. Materials and Methods

We simultaneously collected biometric data for a single participant and data on their environmental context, using an extensive array of sensors. Following data collection, we analyzed this information with machine learning models.

Here is a concise overview of our methodology; a more detailed description of data collection can be found in[11].

2.1 HOLISTIC SENSING

Our holistic sensing has two key parts, environmental and biometrics. The array of environmental sensors consisted of multiple sensors; in this study, we will only consider the CO_2 data measured using a LI-COR LI-850 instrument[16] at a sampling rate of 0.5 Hz. A total of 329 biometric variables have been considered. The EEG data was collected using a Cognionics system[17] at 500 Hz. The ECG, GSR, SpO_2, respiration rate, skin surface temperature, and heart rate were measured using the Cognionics AIM Generation 2 device[18], with a sampling rate of 500 Hz. The physiological readings of the eyes: the diameter of each eye and distance between the pupils were measured using the Tobii Pro Glasses 2[19], with a sampling rate of 100 Hz. The list of all the biometric variables and their purpose is mentioned below:

- Electroencephalography (EEG) is a medical imaging technique that employs sensors to monitor surface brain electrical activity. This method captures electrical signals originating from the brain, which arise from the collective activities of neurons[20]. These electrical signals, measured in voltages, are very small and are measured by placing electrodes on the scalp. In this study, we used an EEG headset equipped with 64 electrodes, adhering to the 10-10 nomenclature system[21]. This device was placed on the scalp, and each electrode was designed to detect minute voltage changes in the brain. The voltage recorded by each electrode was referenced against 'virtual reference', calculated as the average across all channels. The data we obtained took the form of a voltage time series (V) at each electrode. Readings for some electrodes can be distorted with artifacts such as blinks, eye movement, jaw clench, tongue movement, swallowing, neck tension, and head movement which have not been removed.

  The data thus obtained from each electrode can be transformed from the time domain to the frequency domain, and this was done using the Scipy[22] Welch function in the following frequency ranges: Delta (1-3 Hz), theta (4-7 Hz), alpha (8-12 Hz), beta (13-25 Hz) and gamma (25-70 Hz). Therefore, a power spectrum is obtained with the frequency on the X axis in units of Hz and the density of the power spectrum on the Y axis in units (V^2/Hz). The data obtained from each of the 64 electrodes and dividing the frequency into five bands each yield a total of 320 features from the EEG headset alone. The EEG data and code for retrieval of the EEG data and transformation of the EEG data from a voltage time series to power spectral density are uploaded on GitHub and are included in the supplementary materials.

- Electrocardiography (ECG): ECG, or

electrocardiography, is a method employed to measure the electrical activity of the heart. Coordinated electrical impulses in various regions of the heart are essential for maintaining proper blood flow.[23] Analyzing these impulses helps to assess heart rhythm and detect irregularities in heartbeat, as well as determine the rate and strength of these electrical signals[23]. In this study, the ECG measurements were taken in microvolts, with the sensor positioned on the chest surface.

• Galvanic Skin Response (GSR): GSR, also known as skin conductance, is a method that measures the electrical conductivity of the skin. This technique exploits the fact that the sweat glands, which are involuntary, become active in response to emotions such as joy, increasing the conductivity of the skin[24]. However, it is not just emotions that trigger sweating; everyday experiences such as physical labor or exposure to high environmental temperatures can also lead to increased perspiration. In this study, the GSR sensor was positioned on the lower back of the neck. Conductivity is measured in micro Siemens (μSiemens), with higher values indicating a greater sweat response.

• Oxygen Saturation (SpO_2): SpO_2 measurement reflects the proportion of hemoglobin saturated with oxygen compared to hemoglobin not carrying oxygen.[25] For this study, the SpO_2 sensor was placed behind the left ear and the readings were presented as percentages. For example, a reading of 95% SpO_2 indicates that 95% of the hemoglobin in each red blood cell is oxygenated, while the remaining 5% is non-oxygenated.

• Respiration Rate: Indicates the breathing rate per minute, measured with the same device used to measure the GSR.

• Skin Surface Temperature: This measurement reflects the temperature at the skin surface where the sensor is located. In this instance, the sensor was positioned on the right temple of the participant, from their perspective. The temperature is recorded in degrees Celsius (°C).

• Heart rate: Indicates the number of heartbeats per minute measured with the same device used to record SpO_2 data.

• Average Pupil Diameter: This metric represents the mean diameter of the pupils of both eyes, measured in millimeters (mm).

• Pupil Center Distance: This refers to the three-dimensional measurement of the distance between the centers of the pupils, expressed in millimeters (mm).

• Pupil Diameter Disparity: This measures the absolute difference in diameter between the left and right pupil, quantified in millimeters (mm). Sensor placement and biometric suite fitting were carefully managed to minimize physiological responses. All data collected were downsampled to a frequency of 1 Hz, effectively producing one data point per second for each of the 329 variables and the environmental CO_2 measurements.

2.2 DATA COLLECTION

During the COVID-19 pandemic, this study was conducted with a single participant, while plans to involve multiple participants are currently underway. Data collection occurred when the participant rode a bicycle, followed by an electric car equipped with environmental sensors in its trunk, including the CO_2 sensor. Data recording ceased when the bicycle ride ended.

Data gathering took place over three separate days: May 26, June 9, and June 10, 2021. CO_2 measurements were conducted specifically on June 9 and 10, with two trials each day. This study was located in Richardson, Texas.

The cyclist was equipped with a GPS sensor, which tracks longitude, latitude, and altitude. Figure 1, created using MATLAB, displays a street map of the data collection area, displaying the 329 variables and CO_2 measurements recorded simultaneously, resulting in a comprehensive dataset.

Sensor data are not always precise due to measurement artifacts such as biometric sensor movement and may sometimes fail to provide readings, necessitating data cleaning, which leads to occasional gaps in the data. In Figure 1, Trials 1 and 2 exhibit fewer gaps, allowing a nearly continuous path representation, along with the corresponding CO_2 concentrations at various locations. The start and end points of each bicycle ride are marked with an asterisk. The data points collected totaled 710 and 696 for Trials 1 and 2, respectively. In Trials 3 and 4, some data gaps are evident due to cleaning, yielding total data points of 673 and 238, respectively.

The methodology of data collection, downsampling, and subsequent data cleaning resulted in a comprehensive data set comprising 2,317 time-stamped records. This data set was organized into a DataFrame with 330 columns. Of these, 329 columns represent predictor variables, primarily consisting of biometric data, and one column is dedicated to the target variable, which in this case is the CO_2 values. The DataFrame encapsulates these 2,317 data entries, presenting the values of the 330 variables in a time series format. Geographic coordinates (longitude and latitude) along with the corresponding CO_2 values have been made available on GitHub and are included in the supplementary materials.

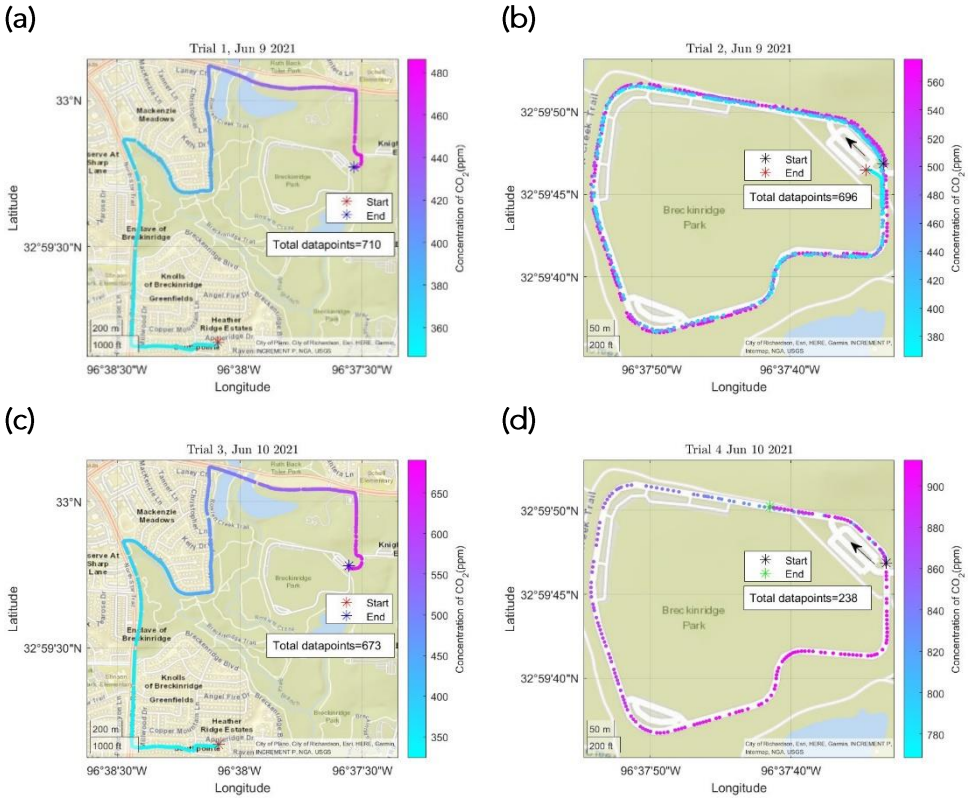

Figure 1: Location and track of bicycle for data collection. **(a)** Track for data collection in Trial 1 on June 9, 2021 with the corresponding CO_2 concentration. **(b)** Track for data collection in Trial 2 on June 9, 2021 with the corresponding CO_2 concentration. Arrow indicates the initial direction of ride in the loop. **(c)** Track for data collection in Trial 3 on June 10, 2021 with the corresponding CO_2 concentration. **(d)** Track for data collection in Trial 4 on June 10, 2021 with the corresponding CO_2 concentration. Arrow indicates the initial direction of ride in the loop.

2.3 DATA ANALYSIS AND MACHINE LEARNING MODEL DEVELOPMENT

Upon constructing the DataFrame, a machine learning algorithm was applied to predict inhaled environmental CO_2 using biometric variables as features. We employed Random Forests[26] for non-linear, multi-dimensional regression, utilizing the scikit-learn[27] Ensemble Random Forest Regressor package. The dataset was split, with 80% used for training the model and the remaining 20% serving as a test set. We evaluated the prediction accuracy by computing Pearson's correlation coefficient between the actual and predicted CO_2 values, where a perfect prediction would result in a coefficient of 1. Additionally, the Root Mean Square Error (RMSE) was calculated to assess the prediction's precision. Qualitative analysis of the actual and predicted values was conducted using Quantile-Quantile plots and Time series plots.

Post-implementation of the Random Forest algorithm, SHAP Values (SHapley Additive explanations)[28], designed for tree-based algorithms[29], were used to rank predictors based on their effectiveness in forecasting the target variable. The SHAP value algorithm was implemented using the TreeExplainer package from the SHAP library.

The top 9 predictor variables were then examined, and a Pearson's correlation coefficient was calculated between these variables and CO_2, resulting in a 10x10

correlation matrix. The corr function from the Pandas library[30,31] was used to investigate potential linear relationships between the variables.

To explore non-linear relationships, mutual information was calculated using the package from scikit-learn[27]. A higher mutual information value indicates a stronger relationship, while a value of zero suggests independence between the variables.

3. Results

The process of estimating the inhaled environmental CO_2 and understanding its connection to the human body through the measured biometrics is broken down into two parts: first, considering all 329 biometrics (or features) of which 320 were EEG variables and 9 were non-EEG (or physiological) variables and second used only the 9 non-EEG variables.

3.1 ANALYSIS USING 329 FEATURES

As mentioned above, the models were trained on 80% of the data set using the scikit-learn random forest algorithm and the remaining 20% was kept as a test. Doing so yielded a very high accuracy in both the training and the testing set, with the value of Pearson's correlation coefficient squared (r^2) between the actual and predicted to be 0.99 and 0.98 respectively. Additionally, the root mean square error (RMSE) between the actual and predicted values was also low, with an RMSE of 9.01 ppm and 25.66 ppm in the training and testing set, respectively. This is much higher than the previous estimation of PM_1 particles that yielded an accuracy of $r^2 = 0.91$ in the test set.[11]

Figure (2a) shows a bar graph of RMSE values with RMSE in the training set in blue, while RMSE in the test set in orange. Figure (2c) shows a scatter plot of the true CO_2 values versus the estimated CO_2 values with the ones in the training set represented by the blue dots and those in the testing set by the orange sign 'x'. It can be seen that most of the points lie at or near the 1:1 dark line, indicating that the difference between the true and estimated values is close most of the time. The points seem to deviate between 700 ppm and 800 ppm where there are fewer data points. The data points then again start to be close to the dark line as the concentration of the data points increases. Figure (2d) shows a quantile-quantile graph of the true values of CO_2 versus the estimated values of CO_2. As can be seen, most of the data points lie close to the 1:1 red line. Again, the quantiles deviate from the red line between the 700 to 800 range, where there are fewer data points available, and start to come closer as the concentration of the data increases. The percentiles of the distribution have also been indicated showing that 75% of the CO_2 values were approximately below the 550 ppm values.

Finally, Figure (2b) shows a SHAP value beeswarm plot of the top 9 features arranged in descending order indicating the features that were the most influential or contributed the most to the prediction of environmental CO_2 inhaled. These SHAP values on the X-axis are in units of the target variables, that is, ppm. The color bar indicates the values of the features with higher values in red and lower values in blue with identical SHAP values for the features stacked vertically in the plot. Depending on how the data are shuffled, the order of the variables might change a bit, but the variables tend to stay the same, especially

in the top seven, where the magnitude of the SHAP values is significantly higher than other values. The average pupil diameter is one of the top most variable useful in predicting the CO_2 values with a higher average pupil diameter lowering the prediction while a lower average pupil diameter increasing the prediction. The dilation of the pupils, as a physiological response, has been correlated with cognitive tasks[32,33]. As data collection is done while the person is cycling, it is expected that the participant is sweating. However, it has been found that inhaling CO_2 causes sweating thereby affecting the GSR values[34]. Similarly, other features such as skin temperature, heart rate, and respiration rate were top features as it was in the estimation of PM_1 particles done previously[11].

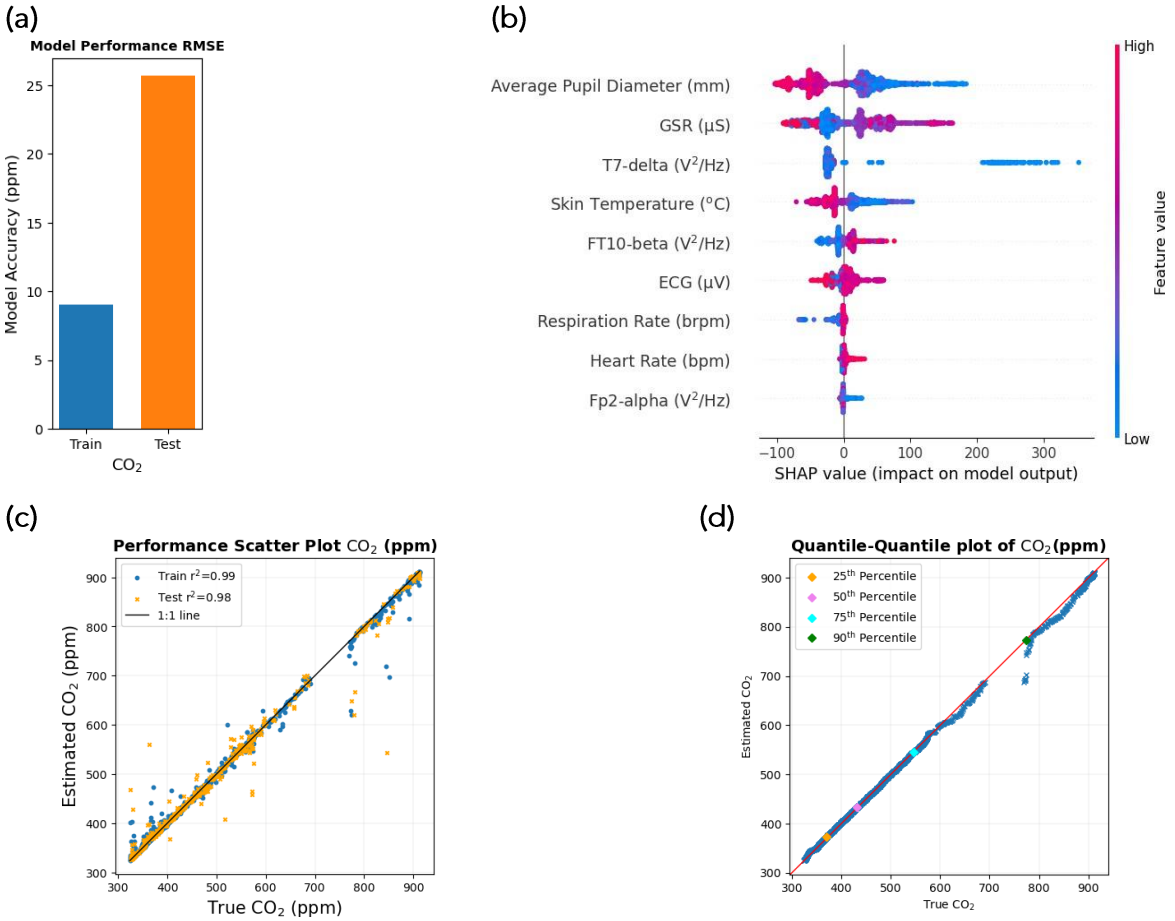

Figure 2: Performance and top 9 feature importance plot for estimating CO_2 using 329 biometric features. **(a)** RMSE between the actual and estimated values of CO_2 in the training and test set. **(b)** Top 9 features in estimating CO_2 identified using SHAP values in a beeswarm plot. **(c)** Scatter plot between the actual and estimated values in the training and the testing set with the corresponding r^2 values. **(d)** Quantile-Quantile plot between the actual and estimated values with the percentiles.

Other variables with the highest contribution in estimating CO_2 included the ones from EEG. The T7 electrode, as suggested by the 10-10 nomenclature system[21], is located in the temporal lobe on the left side of the brain which is involved in speech and short-term memory[35]. The FT10 electrode is located between the frontal and temporal lobe,

located on the right side of the brain. Similarly, the alpha band of the Fp2 electrode which is placed between the frontal and parietal lobes on the right side of the brain seems to have a lesser contribution to the estimation of CO_2 with all variables below, even less contribution with lower SHAP values.

Figure 3 shows a time series of the true CO_2 plotted in solid red with the estimated CO_2 plotted in dotted blue lines for all 4 trials on 2 days of data collection. The shaded background colors differentiate the trials. Trial 1 had among the lowest values of CO_2 while trial 4 had the highest values with CO_2 reaching above 900 ppm. For most parts in the time series, the true values of CO_2 are in close proximity to the estimated values of CO_2.

A 10 by 10 correlation matrix of Pearson's correlation coefficient consisting of the top 9 features and the target variable has been plotted to identify the linear relationship between them and a 10 by 10 matrix of mutual information for the same variables has also been plotted to identify nonlinear relationship. The digits are rounded up to 2 significant figures.

Figure (4a) shows that few of the variables are linearly related. The average diameter of the pupil had a negative correlation with GSR, ECG, heart rate, and vice-versa. GSR had a positive correlation with skin temperature, ECG and vice versa. Regarding the target variable, the strongest linear correlation with CO_2 was found to be with the average diameter of the pupil. Figure (4b) helps in identifying non-linear relationship between the 10 features. The skin temperature had a high mutual information with GSR, ECG, respiration rate, and heart rate. GSR had higher mutual information with ECG, heart rate, respiration rate, and skin temperature. Similarly, with respect to the target variable CO_2 had a high mutual information with respiration rate, heart rate, ECG, skin temperature, GSR which is to be expected, as these variables had a greater contribution to the estimation of CO_2 as shown by SHAP values.

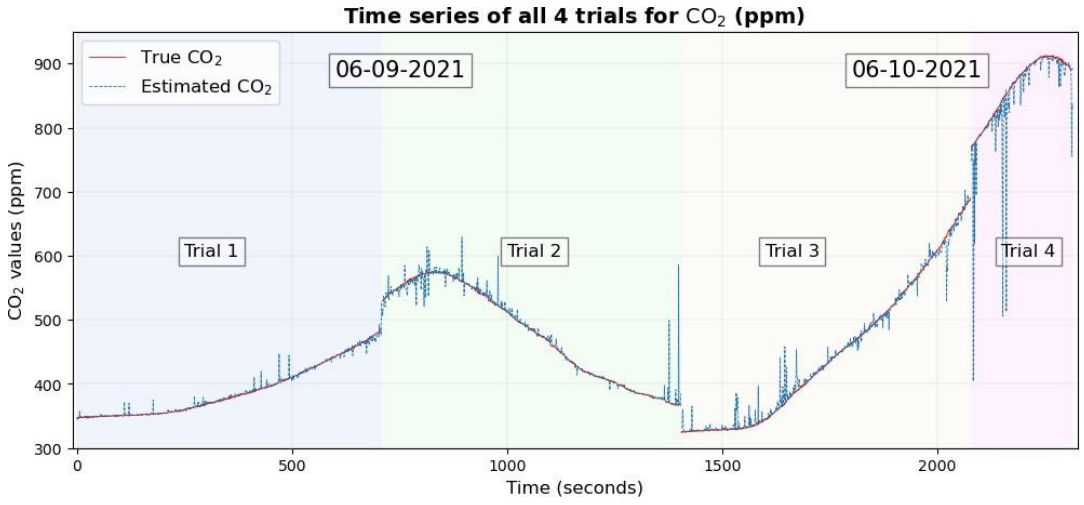

Figure 3: Time series plot of the true CO_2 values with the estimated values of CO_2 overlaid for all the 4 trials of data collected on 2 separate days considering 330 biometric variables

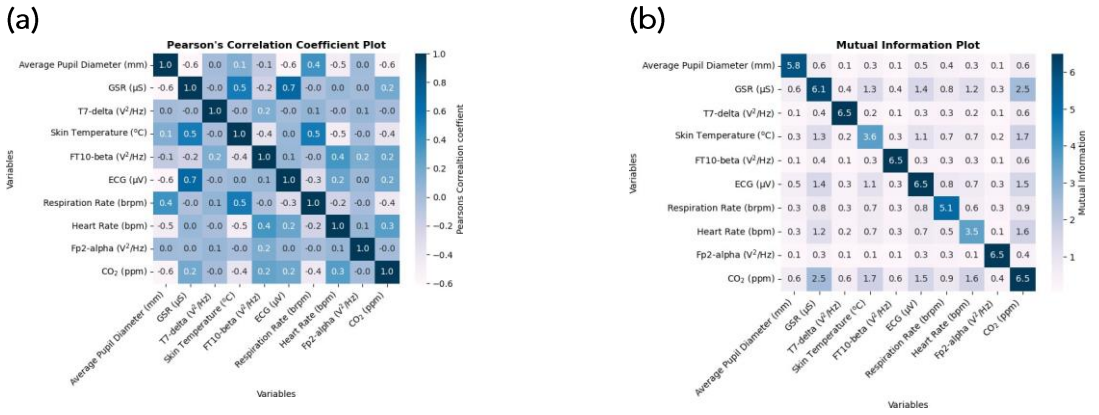

Figure 4: 10 by 10 Pearson's correlation and mutual information matrix **(a)** Pearson's correlation matrix for 10 variables to identify linear relationships. Mutual information matrix for 10 variables to identify linear and non-linear information.

3.2 ANALYSIS USING 9 PHYSIOLOGICAL RESPONSES

Now we consider only the 9 physiological responses (or non-EEG) to estimate CO_2 using the same Random Forest algorithm from scikit-learn with 80% of the data set for training and the remaining 20% of the dataset for testing.

Remarkably, considering only the 9 physiological responses yielded an almost identical and highly accurate result as indicated by Pearson's correlation coefficients of 0.99 and 0.98 between the actual and estimated values of CO_2 in the training and the testing set, respectively. The RMSE between the actual and estimated CO_2 values was also very low, 8.78 ppm and 19.41 ppm in the training and testing set, respectively.

A bar graph of the RMSE in the training set and the testing set is shown in Figure (5a) with the train RMSE in blue and the test RMSE in orange. A scatter plot in Figure (5c) between the true and actual CO_2 values is shown in Figure (5a) with actual values of CO_2 in blue dots and the estimated values in the orange sign 'x'. The plot shows that most of the values are close to the 1:1 dark line with some values deviating between the 700 ppm and 800 ppm just as before with a probable cause being less number of data points.

The quantile-quantile plot in Figure (5c) shows that the true values and the estimated values of CO_2 are very close to each other with values deviating between 700 ppm and 800 ppm with a distribution similar to that of the estimate made using 329 variables.

A SHAP value beeswarm plot in Figure (5b) shows a ranking in descending order of the physiological variables that indicates which of the variables was the most important to predict CO_2. The SHAP value, which in this case is in ppm is close in magnitude for average pupil diameter and GSR thus the ordering of these two variables might change as the data are shuffled. Similarly, the SHAP value for SpO_2, the distance of the pupils, and the absolute value of the difference in the diameter of the pupils are close to each other and the ordering could change when the algorithm is rerun. Just as before, the average pupil diameter, GSR, skin temperature, respiration rate, and heart rate were among the main predictor variables with the immediate next SpO_2 with very low magnitude of SHAP values. Furthermore, the distance of the pupil and the absolute value

of the difference between the pupil diameter have less of a contribution, as indicated by the small magnitude of the SHAP value. Since the SHAP values of SpO_2, the distance of the pupils and the absolute value of the difference in the diameter of the pupils are low, removing these features will have little effect on the result.

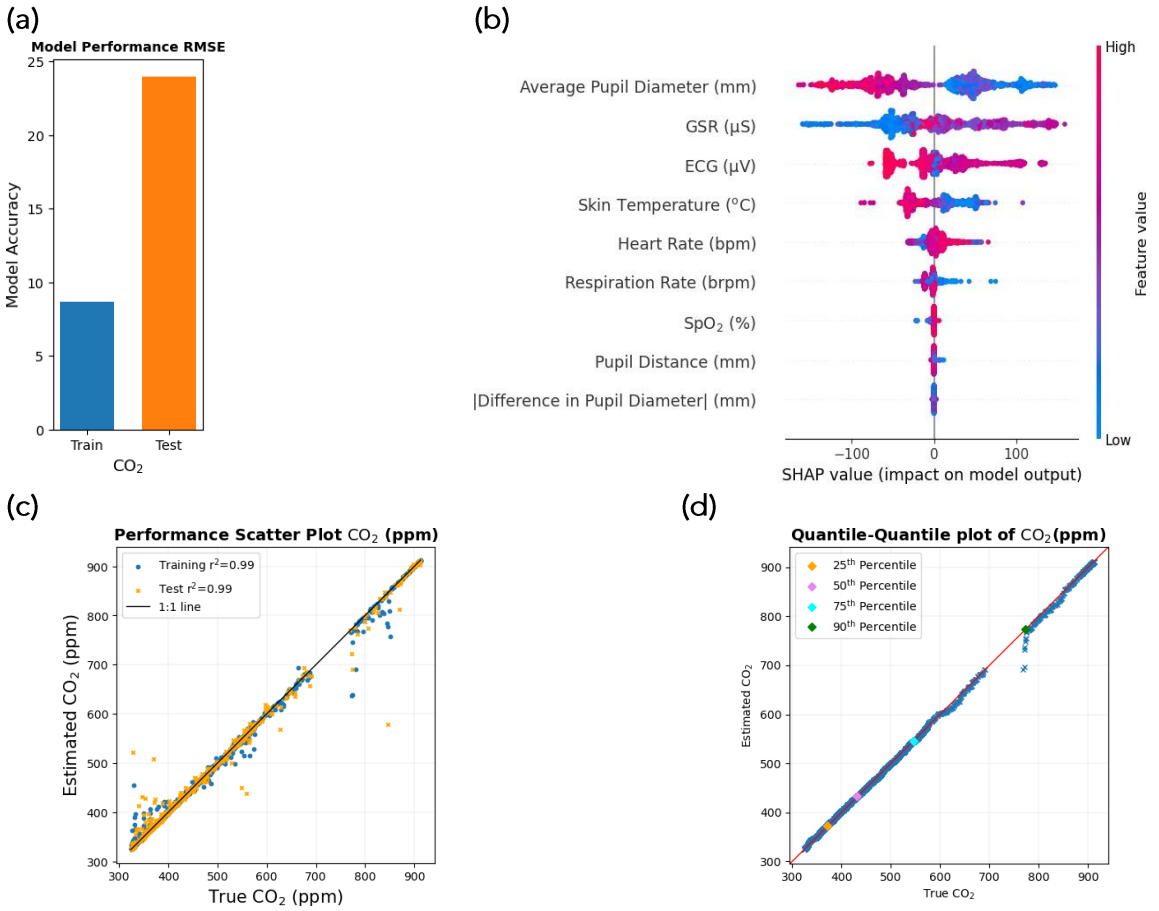

Figure 5: Performance and feature importance plot for estimating CO_2 using 9 biometric features. **(a)** RMSE between the actual and estimated values of CO_2 in the training and the testing set. **(b)** Top 9 features in estimating CO_2 identified using SHAP values in a beeswarm plot. **(c)** Scatter plot between the actual and the estimated values in the training and the testing set with the corresponding r^2 values. **(d)** Quantile-Quantile plot between the actual and the estimated values with the percentiles.

A time series graph of the true CO_2 and the estimated CO_2 of the 9 non-EEG variables is shown in Figure 6. The solid red line represents the true CO_2 whereas the estimated CO_2 is represented by the blue dotted line for the 4 trials in 2 days where the background color indicates the individual trials. Apart from a few points, the estimated CO_2 values are close to the actual CO_2, which was a similar result when considering all the 329 variables.

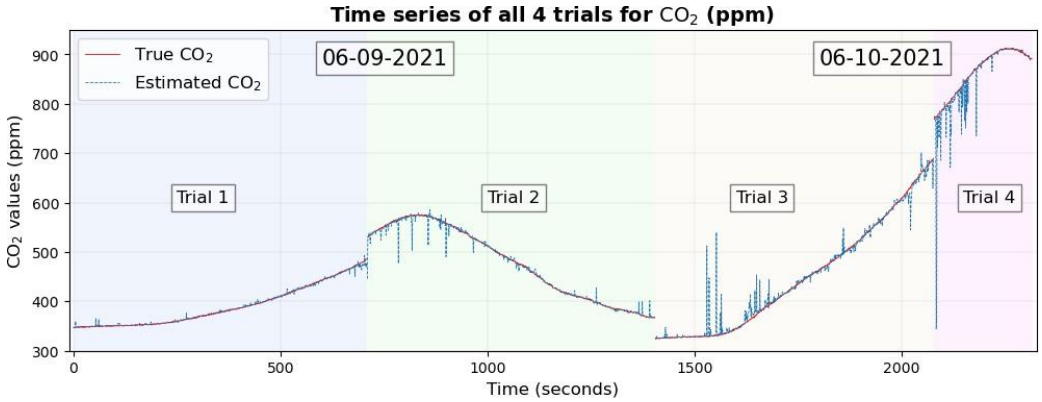

Figure 6: Time series plot of the true CO_2 values with the estimated values of CO_2 overlaid for all the 4 trials of data collected on 2 separate days considering 9 biometric variables

Just as before, a 10 by 10 correlation matrix of Pearson's correlation coefficient and a 10 by 10 mutual information matrix of the 9 physiological responses and CO_2 as shown in Figure 7. A somewhat positive linear correlation is observed between respiration rate and SpO_2. A relationship is expected between respiration rate and SpO_2, since respiration is the way blood receives oxygen.[25] Although there does not appear to be any strong linear correlation of CO_2 with the 9 non-EEG variables expected with that of average pupil diameter as before.

There was greater mutual information between CO_2 and GSR, heart rate, skin temperature, ECG, and average pupil diameter with near zero mutual information between the absolute value of the difference in pupil diameter and the distance of the pupils, which is again expected, as indicated by the magnitude of SHAP values. The absolute value of the difference between the diameters of the pupils seems to have near zero mutual information with most of the other variables. Similarly, the pupil distance also has near-zero mutual information with the rest of the variables, indicating that the variables are almost independent of each other.

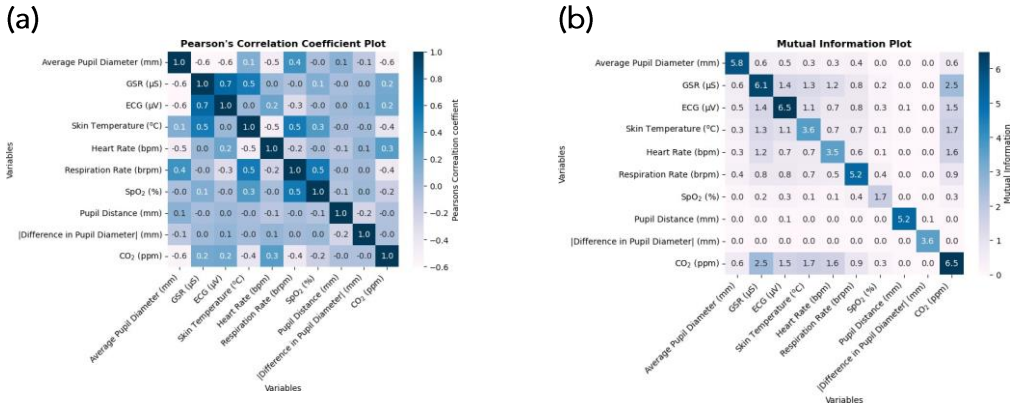

Figure 7: 10 by 10 Pearson's correlation and mutual information matrix (a) Pearson's correlation matrix variables to identify linear relationships. (b) Mutual information matrix for 10 variables to identify linear and non-linear information.

The entire data set and the code for the results are publicly available and are uploaded in GitHub, the link to which is in supplementary materials.

4. Discussion

The results of our study show that the physiological and cognitive responses resulting from inhaling CO_2 and measured using various sensors can be used to predict inhaled CO_2 with precision, as indicated quantitatively by the r^2 and RMSE values between the true and estimated values and qualitatively by the quantile-quantile graph and the time series graph. The predictor ranking indicates that a small set of physiological responses can be used to make the model simpler while also understanding in what way these physiological responses change.

The results obtained from this research show that the methodology previously used to estimate and understand the effects of particulate matter[11] can also be used in inhaled gases such as CO_2. The comprehensive suite that we have used to measure a large number of biometrics can simultaneously test not just the effects of inhaled CO_2 but also the relationship between these variables, as shown by the Pearson correlation matrix to identify linear relationships and the mutual information matrix to capture nonlinear relationships. Moreover, by taking into account a large number of biometrics, we can test not just one but multiple variables that were simultaneously affected due to the inhaled gas.

Although this study showed some promising results, two of the main limitations in this study cannot be overlooked. The first is that the study was conducted with a single participant. To overcome this limitation, measurements were made on multiple days with multiple trials. However, a generalized result can only be obtained by considering a large number of participants. The other limitation is that of artifacts in the EEG signal resulting from various activities such as blinking, tongue movement, jaw clenching, muscle movement, etc. as mentioned before which creates noise in the EEG data. Although there are many algorithms to remove artifacts,[36] considering the activities involved while cycling, which would consist of a combination of artifacts, the removal process can be very challenging. However, the results shown by the study show that the non-EEG variables are good enough to estimate inhaled CO_2.

This work can be extended in various ways. Future work can involve other biometric variables other than the one used and also particulates or gases such as black carbon, carbon monoxide, ozone, etc. A study can also be conducted to test whether the reverse process is possible, for example: can a biometric variable such as skin temperature be predicted with reasonable accuracy using a combination of biometrics and atmospheric compounds such as CO_2. Multidimensional machine learning models can also be used to test whether a biometric variable such as skin temperature can be predicted with other biometrics such as heart rate, breathing rate, and GSR since the mutual information between these variables was found to be high.

5. Conclusion

Using a subset of biometrics, inhaled CO_2 can be estimated using machine learning models with a high fidelity and autonomic responses

can be studied at a small temporal and spatial scale in microenvironments. Although there are a few limitations to this study, there are also mitigation measures with plenty of room for future work. Computational techniques such as machine learning have many practical purposes, which in this case was to understand the effects of inhaled CO_2 on autonomous physiological and cognitive responses of a human body.

Competing Interests:

The authors have declared no competing interests.

Acknowledgements Statement:

We express our gratitude for the support received from the Office of Sponsored Programs, Dean of Natural Sciences and Mathematics, and Chair of the Physics Department at the University of Texas at Dallas.

Funding Statement:

Following grants provide the funding for this study: The US Army (Dense Urban Environment Dosimetry for Actionable Information and Recording Exposure, U.S. Army Medical Research Acquisition Activity, BAA CDMRP Grant Log #BA170483). EPA 16th Annual P3 Awards Grant Number 83996501, entitled Machine Learning Calibrated Low-Cost Sensing. The Texas National Security Network Excellence Fund award for Environmental Sensing Security Sentinels. SOFWERX award for Machine Learning for Robotic Teams.

Supplementary Materials

The data and the code used in this study are publicly available and are available on GitHub: https://github.com/mi3nts/Estimate-CO2 and in Zenodo: https://zenodo.org/records/10184801 (accessed on November 22, 2023)

Ethics Statement

All of the experimental protocols used for this study was approved by The University of Dallas Institutional Review Board, and informed consent was obtained from the participant.

Author Contributions

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

Abbreviations

The abbreviations used in this manuscript are as follows.

ppm	Parts per million
WHO	World Health Organization
EEG	Electroencephalography
GSR	Galvanic Skin Response
ECG	Electrocardiography
PM	Particulate Matter
SpO_2	Blood Oxygen Saturation
RMSE	Root Mean Square Error

References:

[1] NHLBI, "How the lungs work," 2022. https://www.nhlbi.nih.gov/ health/lungs, Last accessed on 2023-05-11.

[2] WHO, "Air pollution," 2023. https://www.who.int/health-topics/ air pollution#tab=tab_1, Last accessed on 2023-05-11.

[3] M. Kampa and E. Castanas, "Human health effects of air pollution," *Environmental Pollution*, vol. 151, no. 2, pp. 362–367, 2008.

[4] J. R. Balmes, J. M. Fine, and D. Sheppard, "Symptomatic bronchoconstriction after short-term inhalation of sulfur dioxide1, 2," *Am Rev Respir Dis*, vol. 136, no. 11171121, pp. 10–1164, 1987.

[5] N. Uysal and R. M. Schapira, "Effects of ozone on lung function and lung diseases," *Current opinion in pulmonary medicine*, vol. 9, no. 2, pp. 144–150, 2003.

[6] J. Kagawa, "Evaluation of biological significance of nitrogen oxides exposure," *The Tokai journal of experimental and clinical medicine*, vol. 10, no. 4, pp. 348–353, 1985.

[7] D. G. Badman and E. R. Jaff´e, "Blood and air pollution; state of knowledge and research needs," *Otolaryngology–Head and Neck Surgery*, vol. 114, no. 2, pp. 205–208, 1996.

[8] K. Ewan and R. Pamphlett, "Increased inorganic mercury in spinal motor neurons following chelating agents.," *Neurotoxicology*, vol. 17, no. 2, pp. 343–349, 1996.

[9] M. Loghman-Adham, "Renal effects of environmental and occupational lead exposure.," *Environmental health perspectives*, vol. 105, no. 9, pp. 928–939, 1997.

[10] D. B. Menzel, "The toxicity of air pollution in experimental animals and humans: the role of oxidative stress," *Toxicology letters*, vol. 72, no. 1-3, pp. 269–277, 1994.

[11] S. Talebi, D. J. Lary, L. O. H. Wijeratne, B. Fernando, T. Lary, M. Lary, J. Sadler, A. Sridhar, J. Waczak, A. Aker, and Y. Zhang, "Decoding physical and cognitive impacts of particulate matter concentrations at ultra-fine scales," *Sensors*, vol. 22, no. 11, 2022.

[12] L. Kajt´ar and L. Herczeg, "Influence of carbon-dioxide concentration on human well-being and intensity of mental work," *QJ Hung. Meteorol. Serv*, vol. 116, no. 2, pp. 145–169, 2012.

[13] U. Satish, M. J. Mendell, K. Shekhar, T. Hotchi, D. Sullivan, S. Streufert, and W. J. Fisk, "Is co2 an indoor pollutant? direct effects of low-to-moderate co2 concentrations on human decision-making performance," *Environmental health perspectives*, vol. 120, no. 12, pp. 1671– 1677, 2012.

[14] X. Zhang, P. Wargocki, and Z. Lian, "Physiological responses during exposure to carbon dioxide and bioeffluents at levels typically occurring indoors," *Indoor air*, vol. 27, no. 1, pp. 65–77, 2017.

[15] P. H. Sechzer, L. D. Egbert, H. W. Linde, D. Y. Cooper, R. D. Dripps, and H. L. Price, "Effect of co2 inhalation on arterial pressure, ecg and plasma catecholamines and 17-oh corticosteroids in normal man," *Journal of Applied Physiology*, vol. 15, no. 3, pp. 454–458, 1960. PMID: 14444401.

[16] LI-COR, "Introduction to the instruments," 2023. https://www.licor.com/env/support/LI-850/topics/description.html#Onlineresources , Last accessed on 2023-08-20

[17] Cognionics, "Products," 2023. https://www.cgxsystems.com/ products, Last

accessed on 2023-08-20.

[18] Cognionics, "Cgx aim physiological monitors," 2023. https://www.cgxsystems.com/auxiliary-input-module-gen2, Last accessed on 2023-08-26.

[19] Tobii, "Tobii pro glasses 2," 2023. https://www.tobii.com/products/discontinued/tobii-pro-glasses-2, Last accessed on 2023-08-27.

[20] M. Soufineyestani, D. Dowling, and A. Khan, "Electroencephalography (eeg) technology applications and available devices," *Applied Sciences*, vol. 10, no. 21, 2020.

[21] J. N. Acharya, A. J. Hani, J. Cheek, P. Thirumala, and T. N. Tsuchida, "American clinical neurophysiology society guideline 2: guidelines for standard electrode position nomenclature," *The Neurodiagnostic Journal*, vol. 56, no. 4, pp. 245–252, 2016.

[22] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[23] John Hopkins Medicine, "Electrocardiogram," 2023. https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/ electrocardiogram, Last accessed on 2023-08-26.

[24] W. B. Albert and T. S. T. Tullis, "Chapter 8 - measuring emotion," in *Measuring the User Experience (Third Edition)* (W. B. Albert and T. S. T. Tullis, eds.), Interactive Technologies, pp. 195–216, Morgan Kaufmann, third edition ed., 2023.

[25] A. Jubran, "Pulse oximetry," *Critical care*, vol. 3, pp. 1–7, 1999.

[26] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[29] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[30] T. pandas development team, "pandas-dev/pandas: Pandas," Feb. 2020.

[31] Wes McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference* (St´efan van der Walt and Jarrod Millman, eds.), pp. 56 – 61, 2010.

[32] S. D. Goldinger and M. H. Papesh, "Pupil dilation reflects the creation and retrieval of memories," *Current directions in psychological science*, vol. 21, no. 2, pp. 90–95, 2012.

[33] P. van der Wel and H. Van Steenbergen, "Pupil dilation as an index of effort in cognitive control tasks: A review," *Psychonomic bulletin & review*, vol. 25, pp. 2005–2015, 2018.

[34] R. W. Bullard, "Effects of carbon dioxide inhalation on sweating," *Journal of applied physiology*, vol. 19, no. 1, pp. 137–141, 1964.

[35] K. H. Jawabri and S. Sharma, *Physiology, Cerebral Cortex Functions*. Treasure Island (FL): StatPearls Publishing, 2023.

[36] X. Jiang, G.-B. Bian, and Z. Tian, "Removal of artifacts from eeg signals: a review," *Sensors*, vol. 19, no. 5, p. 987, 2019.