AWARD NUMBER: W81XWH-18-1-0400

TITLE: Dense Urban Environment Dosimetry for Actionable Information and Recording Exposure (DUE DARE)

PRINCIPAL INVESTIGATOR: Prof. David J. Lary

CONTRACTING ORGANIZATION: University of Texas at Dallas, Richardson, TX

REPORT DATE: JULY 2022

TYPE OF REPORT: FINAL

PREPARED FOR: U.S. Army Medical Research and Development Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JULY 2022	FINAL	30Sep2018 - 29MAR2022

4. TITLE AND SUBTITLE

Dense Urban Environment Dosimetry for Actionable Information and Recording Exposure (DUE DARE)

5a. AWARD NUMBER
W81XWH-18-1-0400

5b. LOG NUMBER
BA170483

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Prof. David J. Lary

E-Mail: david.lary@utdallas.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Texas at
Dallas,
800 W Campbell Rd,
Richardson TX 75080, USA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Development Command

Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In dense urban environments there is currently a lack of accurate actionable information on atmospheric composition (gaseous and particulate) on fine spatial and temporal scales. By simultaneously measuring both the environmental state and the human biometric response we propose a holistic sensing environment and methodology for providing accurate actionable information. A state of the art sensor network involving fixed and mobile sensors using machine learning calibration and uncertainty estimation. Comprehensive wearable biometric sensors are used to characterize the real-time human response to the composition of the air, making the human response an integral part of the sensor network. The holistic sensor network incorporates embedded real time machine learning to increase functionality in providing actionable insights for active human participants.

15. SUBJECT TERMS

None listed.

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					USAMRDC
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**		155	**19b. TELEPHONE NUMBER** *(include area code)*
Unclassified	Unclassified	Unclassified	Unclassified		

TABLE OF CONTENTS

1. **INTRODUCTION:** *Narrative that briefly (one paragraph) describes the subject, purpose and scope of the research.*

Human health and performance are impacted by air pollution and toxic environmental exposure. There is currently a lack of accurate actionable information on atmospheric composition (gaseous and particulate) on fine spatial and temporal scales in dense urban environments. Our primary goal was to provide accurate actionable information through the use of machine learning and multi-scale multi-use sensing. This project's main technical goal was to create a framework for using machine learning to characterize the effects of environmental exposures on human performance. To accomplish this, we made extensive observations of both the environmental state (airborne chemicals and particulates) and the physiological autonomic response as defined by 113 different biometric parameters, the majority of which were measured at 500 Hz, with pupillometric parameters measured at 100 Hz.

2. **KEYWORDS:** *Provide a brief list of keywords (limit to 20 words).*

Toxic environmental exposure.
Human performance.
Biometric measurements.
Machine learning.
Autonomic response.
Airborne particulates.
Dense urban environments.
Actionable insights.
Multi-use multi-scale sensing.

3. **ACCOMPLISHMENTS:** *The PI is reminded that the recipient organization is required to obtain prior written approval from the awarding agency grants official whenever there are significant changes in the project or its direction.*

What were the major goals of the project?

List the major goals of the project as stated in the approved SOW. If the application listed milestones/target dates for important activities or phases of the project, identify these dates and show actual completion dates or the percentage of completion.

The primary technical goal of this study was to create a framework for characterizing how the environment influences human performance by combining holistic multi-use multi-scale sensing with machine learning. To accomplish this, we collected extensive data on the environmental state (airborne chemicals and particulates) as well as the physiological autonomic response, which was characterized by 113 different biometric responses.

The secondary goal was to create, deploy and document a prototype for scalable, reproducible, holistic measurement capabilities that could be easily deployed for multiple applications providing actionable insights, all with a focus on pre-emotive human protection and performance optimization.

What was accomplished under these goals?

For this reporting period describe: 1) major activities; 2) specific objectives; 3) significant results or key outcomes, including major findings, developments, or conclusions (both positive and negative); and/or 4) other achievements. Include a discussion of stated goals not met. Description shall include pertinent data and graphs in sufficient detail to explain any significant results achieved. A succinct description of the methodology used shall be provided. As the project progresses to completion, the emphasis in reporting in this section should shift from reporting activities to reporting accomplishments.

The **major activities** in this project were:

1. Building and deploying a comprehensive **environmental monitoring** system comprised of three distinct components:
    a. Continuous in-situ monitoring of airborne particulates ranging in size from COVID-19 to much larger pollen and airborne mold (0.1 – 40 microns), as well as a suite of gasses.
    b. Mobile in-car reference monitors installed in a zero-emission electric vehicle.
    c. Airborne particulate abundance estimation using remote sensing (from satellites and weather RADARs).
2. Building and deploying a comprehensive **biometric monitoring** system to holistically measure the human autonomic response. Eye tracking glasses measured pupil size and gaze direction at 100 Hz, and a 64 electrode electroencephalography (EEG), electrocardiography (ECG), galvanic skin response (GSR), body temperature, blood oxygen saturation (SpO_2), heart rate (HR), and heart rate variability (HRV) were measured at 500 Hz
3. Comprehensive **machine learning**, both supervised and unsupervised, to build empirical models of human responses to environmental exposures.

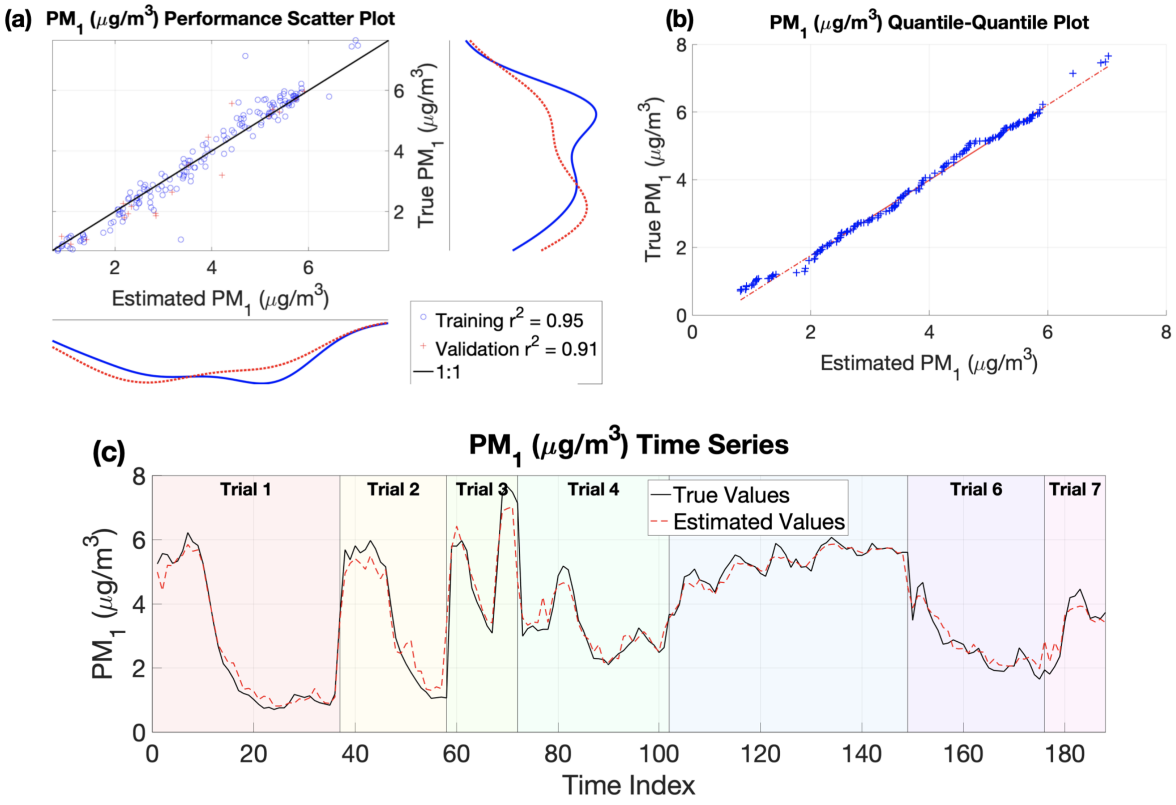
The **specific objectives** in this project were:

1. To create a framework for characterizing how the environment influences human performance by combining holistic multi-use multi-scale sensing with machine learning.
2. To create, deploy and document a prototype for scalable, reproducible, holistic measurement capabilities that could be easily deployed for multiple applications providing actionable insights, all with a focus on pre-emotive human protection and performance optimization.
3. Comprehensive **machine learning**, both supervised and unsupervised, to build empirical models of human responses to environmental exposures.

Some of the significant results include:

1. Using machine learning to accurately model the autonomic pupillary response light intensity of various wavelengths. The human body has numerous autonomic responses. Changes in light intensity, for example, cause pupil dilation to change. Traditionally, the pupil size formulae based on luminance have had very poor accuracy. We used machine learning to investigate the multivariate non-linear autonomic response of pupil dilation as a function of a comprehensive set of over 400 environmental parameters in this project, resulting in the provision of the most accurate model to date. A multivariate non-linear non-parametric supervised regression algorithm with an ensemble of regression trees receiving input from both spectral and biometric data was used in the objectively optimized empirical machine learning models. The models for predicting the participants' pupil diameters from the input data had a fidelity of at least 96.9% for both the training and independent validation data sets. The most important inputs were light levels (irradiance) at wavelengths near 562 nm. This corresponds to the maximum absorbance of the retina's long-wave photosensitive cones, which have a maximum absorbance of 562.8±4.7 nm.

2. Discovering that, just as humans have a rapid autonomic response to light intensity, for example, if a human gaze moves to a bright light in a few milliseconds, pupil dilation autonomically decreases, there is also a rapid discernible autonomic response when airborne particulates are inhaled. This autonomic response can be calibrated to determine how many particles per cm^3 of a specific size are inhaled. The figure below is from our publication available at https://doi.org/10.3390/s22114240.



A scatter plot of measured PM_1 abundance versus that inferred solely from biometric measurements using machine learning is shown in panel (a). A perfect fit is indicated by the black 1:1 line. Training data are represented by blue circles, while validation data are represented by red pluses. Panel (b) depicts the corresponding quantile-quantile plot of measured PM_1 versus that inferred using machine learning solely from biometric measurements. A perfect y = x line would result from identical true and predicted distributions. Panel (c) depicts a time-series plot of measured PM_1 values (solid black line) and those inferred solely from biometric measurements using machine learning across seven separate data collections (dashed red line). The background color represents the trial number for each time period. Trials 1-3 were collected on May 26th, 2021; trials 4-5 on June 9th, 2021; and trials 6-7 on June 10th, 2021.

3. Developing an algorithm for Unsupervised Blink Detection Using Eye Aspect Ratio Values. The eyes provide access to underlying physical and cognitive processes. Although pupil size has been extensively studied, blinking is a less studied but potentially informative factor. Blink detection techniques are far less common than eye-tracking and pupil size estimation tools due to their novelty. We built a new unsupervised machine learning blink detection strategy that makes use of existing eye-tracking technology. The method is compared to two other methods. For blink detection, all three algorithms use eye aspect ratio values. Accurate and rapid blink detection adds to existing eye-tracking research and could provide a new informative index of physical and mental health.
4. Developing an algorithm for Data-Driven EEG Band Discovery with Decision Trees. Electroencephalography (EEG) is a brain imaging technique that involves placing electrodes on the scalp. EEG signals are commonly divided into four frequency bands: delta, theta, alpha, and beta. While these bands have been demonstrated to be useful for characterizing various brain states, their utility as a one-size-fits-all analysis tool is unknown. We developed an algorithm that provides an objective strategy for identifying optimal EEG bands based on signal power spectra. A two-step data-driven methodology for objectively determining the best EEG bands for a given dataset is presented. First, for a predetermined number of bands, a decision tree is used to estimate the optimal frequency band boundaries for reproducing the signal's power spectrum. An Akaike Information Criterion (AIC)-inspired quality score that balances goodness-of-fit with a small band count is then used to determine the optimal number of bands. By identifying bands that outperformed the more commonly used band boundaries by a factor of two, this data-driven approach resulted in a better characterization of the underlying power spectrum. Key spectral components were also isolated in dedicated frequency bands. The proposed method offers a fully automated and adaptable approach to capturing key signal components and potentially discovering new brain activity indices.

5. Developing a machine learning algorithm for the calibration of low-cost Airborne Particulate Sensors. Airborne particulates are particularly important due to their effects on human health as well as their roles in both atmospheric radiative transfer and atmospheric chemistry. Environmental agencies typically collect data on airborne particulates using expensive instruments. The number of sensors that can be deployed is limited due to the high cost of the instruments typically used by environmental agencies. We show how machine learning can be used to effectively calibrate lower-cost optical particle counters in this study. It is critical for this calibration to take measurements of the atmospheric pressure, humidity, and temperature.
6. Developing a machine learning algorithm for the calibration of low-cost light sensors. Sunlight incident on the Earth's atmosphere is necessary for life, and it drives a variety of photochemical and environmental processes, such as radiative heating of the atmosphere. We describe and apply a physical methodology for providing wavelength resolved irradiance spectra with a resolution of 1 nm between 360-780 nm by calibrating against a reference sensor using machine learning using an ensemble of very low-cost sensors (with a total cost of $20, less than 0.5% of the cost of the reference sensor). These low-cost sensor ensembles are calibrated using machine learning and can accurately reproduce the observations made by a NIST calibrated reference instrument (Konica Minolta CL-500A, which costs around $6,000). $R^2 > 0.99$ has been optimized for the correlation coefficient between the reference sensor and the calibrated low-cost sensor ensemble. Both the circuits and the code used have been made public. We can distribute a large number of low-cost sensors in a neighborhood scale area by accurately calibrating the low-cost sensors. It provides unprecedented spatial and temporal insights into the micro-scale variability of wavelength resolved irradiance, which is useful for air quality, environmental, and agricultural applications.

7. Physical Quantification of the Interactions Between Environment, Physiology, and Human Performance. Characterizing key physical interactions between the human body and its surroundings has numerous important applications in public health, preventative healthcare, city planning, sports medicine, aviation, and other fields. The complexity of these multifaceted interactions, on the other hand, makes physical first principles approaches difficult. We authored a paper in which we describe a data-driven experimental paradigm that combines holistic physical sensing with a variety of computational tools to generate empirical machine learning models that quantify the interactions between environment, human physiology, and performance. This paradigm's two main outputs are 1) high-fidelity predictive models and 2) objective evaluation of predictor variable impacts on target variables. Particulate concentrations, for example, were accurately inferred from biometric observations alone in one case study using an empirical machine learning model. Following that, evaluation model predictors revealed that body temperature was the best predictor of particulate concentrations. This adaptable paradigm is applied in a variety of contexts to provide practical insights into the complex, interconnected dynamics of environment, physiology, and human performance.

What opportunities for training and professional development has the project provided?
If the project was not intended to provide training and professional development opportunities or there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe opportunities for training and professional development provided to anyone who worked on the project or anyone who was involved in the activities supported by the project. "Training" activities are those in which individuals with advanced professional skills and experience assist others in attaining greater proficiency. Training activities may include, for example, courses or one-on-one work with a mentor. "Professional development" activities result in increased knowledge or skill in one's area of expertise and may include workshops, conferences, seminars, study groups, and individual study. Include participation in conferences, workshops, and seminars not listed under major activities.

A total of 90 students (names listed below) were involved with this project with extensive one on one mentorship throughout:
12 graduate students.
69 undergraduate students.
9 high school students.
6 PhD dissertations, 9 publications with another 1 in progress.

Graduate Students

1. Lakitha Omal Harindha Wijerante (thirteen publications, PhD defended)
2. Gebreab K. Zwedie (fourteen publications, PhD defended)
3. Xun Liu (two publications, PhD defended)
4. Xiaohe Yu (three publications, PhD defended)
5. Yichao Zhang (two publication, publication, PhD defended)
6. Shawhin Talebi (eight publications accepted, won Deans' poster award, PhD defended)
7. Adam Aker
8. John Waczak
9. Bharana Fernando
10. Ruwali Shisir
11. Prabuddha Madusanka
12. Mazhar Iqbal

Undergraduate Students

13. Daniel Kiv, awarded Undergraduate Research Scholar Awards for "Low cost air quality internet of things."
14. Aaron Barbosa
15. Berkley Shofner (Clark Research Program)
16. John Charles Sadler
17. Arjun Sridhar
18. Nikhil Prassan Narvekar
19. Julia Boah Kim
20. Giakhanh Huu Hoang

Undergraduate Senior Design Students Fall 2019 - UTD MINTS Biometrics Project

21. Ritika Shrivastava
22. Shlok Kothari
23. Amna Ali
24. Jacqueline Solis

Undergraduate Senior Design Students Fall 2019 - UTD MINTS Air Quality Project

30. Kameron Noorbakhsh
31. Nicholas Steele
32. Nikhil Nair
33. Jake Schroder
34. Benjamin Hogan

Undergraduate Senior Design Students Spring 2020 - UTD MINTS Biometrics Project

35. Nikhil Nannapaneni
36. Albin Mathew
37. William Hood
38. Marco Myers
39. Akito Ito
40. Bikram Singh

Undergraduate Senior Design Students Spring 2020 – UTD MINTS Air Quality Project

41. Jacob Scheller
42. Jonah Duncan
43. Getenet Demsie
44. Nathan Nguyen

Undergraduate Senior Design Students Fall 2020 - UTD MINTS Biometrics Project

45. Vihasreddy Gowreddy
46. Madhav Mehta
47. Ryan Rahman
48. Arjun Sridhar
49. Rohit Shenoy

Undergraduate Senior Design Students Fall 2020 - UTD-MINTS Air Quality Project

50. Bryanth Fung
51. Keigo Ma
52. Robert Wu
53. Kangzhi Zhao

Undergraduate Senior Design Students Spring 2021 - UTD MINTS Biometrics Project

54. Rami Jaber
55. Jesse Ladyman
56. Cristian Xavier Garces
57. Bradley Krakar

Undergraduate Senior Design Students Spring 2021 – UTD MINTS Air Quality Project

58. Sidney Evans
59. Kevin Flores
60. Fawaz Khurram
61. Veronica Ramirez
62. Daniel Yustana

Undergraduate Senior Design Students Fall 2021 - UTD MINTS Biometrics Project

63. Rolando Martinez
64. Omar Luna
65. Michael Lee
66. Sopuruchi Chisom
67. Nikhil John

Undergraduate Senior Design Students Fall 2021 – UTD MINTS Air Quality Project

68. Keshav Dhamanwala
69. Aditya Agrawal
70. Tommy Symalla
71. Dien Tran
72. Michael Villordon

Undergraduate Senior Design Students Spring 2022 – UTD MINTS Air Quality Project

73. Michael Spencer
74. Rishi Chandna
75. Anthony Maranto
76. Usaid Malik

Undergraduate Senior Design Students Spring 2022 – UTD MINTS Air Quality Project

77. Basil El-Hindi
78. George Yi
79. Eric Zhang
80. Trent Haines
81. Noah Barber

High School Students

5 during summer 2019
4 during summer 2021

How were the results disseminated to communities of interest?
If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe how the results were disseminated to communities of interest. Include any outreach activities that were undertaken to reach members of communities who are not usually aware of these project activities, for the purpose of enhancing public understanding and increasing interest in learning and careers in science, technology, and the humanities.

Publications:

1. Talebi, S., Lary, D. J., Wijerante, L. O. H., & Lary, T. (2019). Modeling Autonomic Pupillary Responses from External Stimuli using Machine Learning. Biomedical Journal of Scientific & Technical Research, 20(3), 14999–15009.
2. Wijeratne, L. O. H., Kiv, D. R., Aker, A. R., Talebi, S., & Lary, D. J. (2020). Using Machine Learning for the Calibration of Airborne Particulate Sensors. Sensors, 20(1), 99.
3. Lary, D. J., Schaefer, D., Waczak, J., Aker, A., Barbosa, A., Wijeratne, L. O. H., Talebi, S., Fernando, B., Sadler, J., Lary, T., & others. (2021). Autonomous Learning of New Environments with a Robotic Team Employing Hyper-Spectral Remote Sensing, Comprehensive In-Situ Sensing and Machine Learning. Sensors, 21(6), 2240. Publication from SOFWERX follow on project, also led to NASA Tech Briefs Q&A: Team of Robots Maps Composition of an Environment. NASA Tech Briefs, 45(6) https://www.techbriefs.com/component/content/article/tb/pub/techbriefs/electronics-and-computers/39254. The government released video is available at https://youtu.be/-VB3og5qmG0
4. Zhang Y, Wijeratne LOH, Talebi S, Lary DJ. Machine Learning for Light Sensor Calibration. Sensors. 2021; 21(18):6259. https://doi.org/10.3390/s21186259 and https://www.mdpi.com/1424-8220/21/18/6259/htm
5. Yu X, Lary DJ, Simmons CS. PM2.5 Modeling and Historical Reconstruction over the Continental USA Utilizing GOES-16 AOD. Remote Sensing. 2021; 13(23):4788. https://doi.org/10.3390/rs13234788 and https://www.mdpi.com/2072-4292/13/23/4788/htm

6. Yu X, Lary DJ, Simmons CS, Wijeratne LOH. High Spatial-Temporal PM2.5 Modeling Utilizing Next Generation Weather Radar (NEXRAD) as a Supplementary Weather Source. Remote Sensing. 2022; 14(3):495. https://doi.org/10.3390/rs14030495 and https://www.mdpi.com/2072-4292/14/3/495/htm
7. Fernando, B.A.; Sridhar, A.; Talebi, S.; Waczak, J.; Lary, D.J. Unsupervised Blink Detection Using Eye Aspect Ratio Values. *Preprints* **2022**, 2022030200 (doi: 10.20944/preprints202203.0200.v1) https://www.preprints.org/manuscript/202203.0200/v1.
8. Talebi, Shawhin, John Waczak, Bharana A. Fernando, Arjun Sridhar, and David J. Lary. 2022. "Data-Driven EEG Band Discovery with Decision Trees" *Sensors* 22, no. 8: 3048. https://doi.org/10.3390/s22083048
9. Shawhin Talebi, David Lary, Lakitha Wijeratne et al. Decoding Physical and Cognitive Impacts of PM Concentrations at Ultra-fine Scales, 29 March 2022, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-1499191/v1]
10. PhD Dissertation in Physics, Gebreab Zewdie, Using a Comprehensive Characterization of the Physical Environment and Machine Learning to Forecast the Abundance of Airborne Pollen, May 2019, The University of Texas at Dallas.
11. PhD Dissertation in Physics, Xun Liu, Physical Studies of Airborne Pollen and Particulates Utilizing Machine Learning, December 2019, The University of Texas at Dallas.
12. PhD Dissertation in Physics, Lakitha Omal Harindha Wijeratne, Coupling Physical Measurement With Machine Learning for Holistic Environmental Sensing, May 2021, The University of Texas at Dallas.
13. PhD Dissertation in GIS, Xiaohe Yu, Cloud Detection and PM2.5 Estimation Using Machine Learning, October 2021, The University of Texas at Dallas.
14. PhD Dissertation in Physics, Yichao Zhang, Providing Wavelength Resolved Irradiance Measurements by Using Machine Learning, May 2022, The University of Texas at Dallas.
15. PhD Dissertation in Physics, Shawhin Talebi, Physical Quantification of the Interactions Between Environment, Physiology, and Human Performance, May 2022, The University of Texas at Dallas.

Public Presentations so far

1. Lary, D. J., Talebi, S., Wijeratne, L., Aker, A., Yu, X., Zhang, Y., Lary, T., Waczak, J., Fernando, B., & Balagopal, G. (2020). Cognitive Performance and the Environment. Virtual Frontiers.
2. Lary, D. J. (2020). Fun of Physics: Physics in Service of Society. UT Dallas, Fun of Physics Seminar Series.
3. Lary, D. J., Talebi, S., Wijeratne, L., Aker, A., Yu, X., Zhang, Y., Lary, T., Waczak, J., Fernando, B., & Balagopal, G. (2020). Physics in Service of Society. UT Dallas, Physics Departmental Seminar.
4. Lary, D. J., Talebi, S., Wijeratne, L., Aker, A., Yu, X., Zhang, Y., Lary, T., Waczak, J., Fernando, B., & Balagopal, G. (2020). Machine Learning and Holistic Sensing for Societal Benefit. UT Dallas, Bioengineering Departmental Seminar.
5. Lary, D. J. (2021). Shared Air DFW Community Air Monitoring Network. Air North Texas Coalition.
6. Lary, D. J. (2021). Good Health and Well Being: Machine Learning and Holistic Sensing for Societal Benefit. Regional Center of Expertise on Education for Sustainable Development (RCE North Texas), 2021 Virtual Annual Summit – United Nations University.
7. Wijeratne, L., Kiv, D., Aker, A., Balagopa, G., & Lary, D. J. (2021). Machine Learning Calibrated Low-Cost Sensing. EPA P3 (People Prosperity Planet) National Student Design Expo, 2021.
8. Lary, D.J., Sensing in Service of Society, Research 411 Talk Show, March 30, 2022

Websites

1. Live environmental data. We are committed to open data and open source. All our environmental data is available online in real-time as a live map at https://www.sharedairdfw.com. The map shows in one place our sensor data, the EPA data, weather radar data, wind data, pollution sources and satellite data. The legend in the top left allows you to turn on and off the various data sources. This approach is now being rolled out at scale leading too many follow-on partnerships. The map is used by many community groups and Dallas County.
2. All the sensor designs, sensor code, portal code and other biometric analysis software has been made open source and is already available in **113 packages** available at https://github.com/mi3nts.
3. MINTS-AI: Multi-Scale Integrated Intelligent Interactive Sensing for Actionable Insights https://mints.utdallas.edu

4. **IMPACT:** D*escribe distinctive contributions, major accomplishments, innovations, successes, or any change in practice or behavior that has come about as a result of the project relative to:*

What was the impact on the development of the principal discipline(s) of the project?
If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe how findings, results, techniques that were developed or extended, or other products from the project made an impact or are likely to make an impact on the base of knowledge, theory, and research in the principal disciplinary field(s) of the project. Summarize using language that an intelligent lay audience can understand (Scientific American style).

The comprehensive multi-scale multi-use sensing in this project catalyzed three synergistic projects one with USSOCOM and two with SOFWERX.

1. The biometric suite developed for this project was used in a USSOCOM POTFF project led by the USUHS. It was a Pilot Study of Physiological Performance Predictors in High-Stress, Live-Fire Scenarios. Autonomic responses, such as the stress response, are physiological mechanisms that are designed to keep us safe. They offer insight into the underlying cognitive and physiological processes. Processing novel life-threatening stimuli depletes cognitive resources, which can impair decision-making and performance. A deeper and more detailed understanding of autonomic physiological responses in the context of high-stress live-fire scenarios opens the door to interventions that improve performance while also saving lives. We used a comprehensive picture of the participants' physical and cognitive status in the context of hyper-realistic live-fire training scenarios to predict performance in this pilot study. The high-dimensional space of biometric markers lends itself to the development of objectively optimized empirical models of performance using machine learning.

2. The low-cost environmental sensing suite developed for this project was used in two SOFWERX projects for the Autonomous Learning of New Environments with a Robotic Team Employing Hyper-Spectral Remote Sensing, Comprehensive In-Situ Sensing and Machine Learning. This project demonstrated an autonomous robotic team that can quickly learn the characteristics of new environments. The adaptable paradigm is applicable to satellite calibration/validation and the development of new remote sensing data products, and it is easily scalable to multi-robot, multi-sensor autonomous teams. A case study for rapid characterization of the aquatic environment is described; in just a few minutes, we collected thousands of training data points. This training data enabled our machine learning algorithms to rapidly learn by example and provide wide-area maps of the environment's composition. Along with these larger autonomous robots, two smaller robots that can be deployed by a single person (a walking robot and a robotic hover-board) were deployed, revealing significant small scale spatial variability. The Autonomous Robotic Team was documented in the following:
    a. Two Department of Defense demonstration Government release Videos:
        i. https://youtu.be/-VB3og5qmG0 and
        ii. https://youtu.be/_X8cKNC7Hn0
    b. A NASA Tech Brief: https://www.techbriefs.com/component/content/article/tb/pub/features/qa/39254
    c. Journal Article: https://www.mdpi.com/1424-8220/21/6/2240
    d. Dallas Morning News (11:36 AM on Jul 6, 2022) https://www.dallasnews.com/news/2022/07/06/environmental-cleanup-robots-being-trained-at-ut-dallas-to-tackle-hurricanes-oil-spills/

What was the impact on society beyond science and technology?
If there is nothing significant to report during this reporting period, state "Nothing to Report."

Describe how results from the project made an impact, or are likely to make an impact, beyond the bounds of science, engineering, and the academic world on areas such as:

- *improving public knowledge, attitudes, skills, and abilities;*
- *changing behavior, practices, decision making, policies (including regulatory policies), or social actions; or*
- *improving social, economic, civic, or environmental conditions.*

This study's sensor network of sensors has been and continues to be expanded through collaboration with a variety of community groups and cities. One of these led to the closure of an illegal dump called "Shingle Mountain" located in the first freedman's community in Joppa, Dallas TX. This then became the subject of a TV documentary which can be seen at https://www.bet.com/episodes/8198nh/disrupt-dismantle-shingle-mountain-season-1-ep-1.

5. **CHANGES/PROBLEMS:** *The PD/PI is reminded that the recipient organization is required to obtain prior written approval from the awarding agency grants official whenever there are significant changes in the project or its direction. If not previously reported in writing, provide the following additional information or state, "Nothing to Report," if applicable:*

The only changes that occurred were due to COVID-19. This led to some delays which were addressed with a no-cost extension.

6. SPECIAL REPORTING REQUIREMENTS

QUAD CHARTS: *If applicable, the Quad Chart (available on https://www.usamraa.army.mil/Pages/Resources.aspx) should be updated and submitted with attachments.*

Dense Urban Environment Dosimetry for Actionable Information and Recording Exposure (DUE DARE)

BA170483

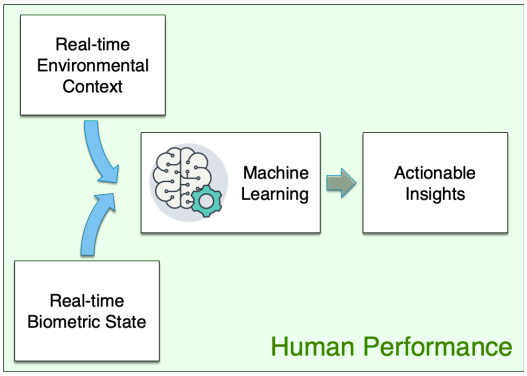
PI: Prof. David J. Lary **Org:** University of Texas at Dallas **Award Amount:** $558,235

Study Aims

In dense urban environments there is currently a lack of accurate actionable information on atmospheric composition (gaseous and particulate) on fine spatial and temporal scales. By simultaneously measuring both the environmental state and the human biometric response we propose a holistic sensing environment and methodology for providing accurate actionable information.

Approach

A state of the art sensor network involving fixed and mobile sensors using machine learning calibration and uncertainty estimation. Comprehensive wearable biometric sensors are used to characterize the real-time human response to the composition of the air, making the human response an integral part of the sensor network. The holistic sensor network incorporates embedded real time machine learning to increase functionality in providing actionable insights for active human participants.



Timeline and Cost

Activities CY	2018	2019	2020	2021	2022
Sensor Acquisition & Calibration – **Milestones**: Low cost sensor calibration/Publication/IRB/HRPO					
Electric Vehicle Integration **Milestones**: Test Survey					
Measurement Campaigns – **Milestones**: Deployment of low cost sensors & Surveys					
Machine Learning Analysis – **Milestones**: Publication/Final Report/Fort Detrick presentation					
Estimated Budget	$200k	$300k	balance		

Updated: UT Dallas, March, 2022

Goals/Milestones

CY18 Goals – Sensor Acquisition & Calibration
- Sensor acquisition

CY19 Goals – Electric Vehicle Integration & Measurement Campaigns
- Low-cost sensor calibration and deployment
- Vehicle sensor suite training
- Vehicle sensor suite testing
- Publication/Conference Presentation
- Integration of vehicle sensors into sensor pod
- Integration of sensor pod into car

CY20 Goals – Machine Learning Analysis
- Survey with participants
- Machine learning analysis linking biometric responses to environmental triggers

CY21 Goals – Machine Learning Analysis
- Survey with participants
- Machine learning analysis linking biometric responses to environmental triggers

CY22 Goals – Completion
- Survey remaining participants
- Analysis linking biometric responses to environmental triggers. Final report & publications.

7. **APPENDICES:** *Attach all appendices that contain information that supplements, clarifies or supports the text. Examples include original copies of journal articles, reprints of manuscripts and abstracts, a curriculum vitae, patent applications, study questionnaires, and surveys, etc.*

Research Article

BIOMEDICAL Journal of Scientific & Technical Research

ISSN: 2574 -1241

DOI: 10.26717/BJSTR.2019.20.003446

Modeling Autonomic Pupillary Responses from External Stimuli Using Machine Learning

Shawhin Talebi*, David J. Lary, Lakitha O.H. Wijeratne and Tatiana Lary

William B. Hanson Center for Space Sciences, Department of Physics, University of Texas at Dallas, USA

***Corresponding author's:** Shawhin Talebi, William B. Hanson Center for Space Sciences, Department of Physics, University of Texas at Dallas, 800 W Campbell Rd, Richardson TX 75080, USA

ARTICLE INFO

Received: July 30, 2019

Published: August 08, 2019

Citation: Shawhin Talebi, David J. Lary, Lakitha O.H. Wijeratne, Tatiana Lary. Modeling Autonomic Pupillary Responses from External Stimuli using Machine Learning. Biomed J Sci & Tech Res 20(3)-2019. BJSTR. MS.ID.003446.

Abstract

The human body exhibits a variety of autonomic responses. For example, changing light intensity provokes a change in the pupil dilation. In the past, formulae for pupil size based on luminance have been derived using traditional empirical approaches. In this paper, we present a different approach to a similar task by using machine learning to examine the multivariate non-linear autonomic response of pupil dilation as a function of a comprehensive suite of more than four hundred environmental parameters leading to the provision of quantitative empirical models. The objectively optimized empirical machine learning models use a multivariate non-linear non-parametric supervised regression algorithm employing an ensemble of regression trees which receive input data from both spectral and biometric data. The models for predicting the participant's pupil diameters from the input data had a fidelity of at least 96.9% for both the training and independent validation data sets. The most important inputs were the light levels (irradiance) of the wavelengths near 562 nm. This coincides with the peak sensitivity of the long-wave photosensitive cones in the retina, which exhibit a maximum absorbance around λ_{max} = 562.8 ± 4.7 nm.

Introduction

This study is part of a broader investigation into the role of the environment in influencing human physical and cognitive performance. The main purpose of this paper is to provide a baseline which accurately describes how changing illuminace affects pupil dilation, so that when emotional or cognitive factors are also involved, we can start to discern the relative roles of illumnance and cognitive load in affecting the pupil dilation [1-3]. The ranking of the importance of the predictor variables used in our empirical machine learning models provides a useful metric of which variables are the key drivers, providing us with valuable insights. The Autonomic Nervous System (ANT) is responsible for changes in pupil dilation. The changes in pupil dilation may occur due to changing light intensity, cognitive load and emotional load [4]. While the light intensity allows an immediate response at the retinal level, an emotional and especially cognitive response, require some higher level processing. So, when the visual input is sent from the eye to the visual cortex via the optic nerve, it first goes through the thalamus. If at this point an imminent threat is detected, it responds mobilizing the body for a 'fight or flight' response, which is then reflected in the changes in the pupil size. As the visual information is relayed to the visual center of the brain in the occipital lobe, it is further sent for processing via various routes to different parts of the brain. In a fast paced changing environment, executive function in the prefrontal lobes make decisions in a fraction of a second. This process also effects changes in pupil dilation. Some areas of the brain involved in the processing of cognitive and emotional load are deep seated structures and can only be observed by expensive equipment such as fMRI in an artificial lab setting. So, part of the question we are starting to address in this study is how can we tell the difference to which stimuli the pupil is responding? This study begins to answer this question using non-invasive methods that can be used in a natural setting by providing a methodology to accurately model the change in pupil size as a function of key environmental variables, so that when other changes are also occurring simultaneously (such as emotional and cognitive load) we can start to examine how these factors modify the pupil dilation response that occurs.

In addition to changes in pupil dilation, other autonomic responses include changes in heart rate variability, galvanic skin

response (or sweating), and core temperature [5-7]. Each of these responses are influenced by variables such as cognitive load [8-11], age [12], pain level [13], and emotional state [14]. In several previous studies formulae for pupil size utilized a single variable, luminance [15-19]. A major shortcoming of these models is their lack of generality. This is illustrated in Figure 1, where the true pupil diameter is plotted against the estimated pupil diameter provided by each of the models enumerated in the legend. There is a clear contrast between the diffuse *cloud* of data points from previous model predictions and the high fidelity predictions of the machine learning model developed here, shown by the green (training points) and the red (independent validation points) in the foreground. Of the five previous models, Holladay's formula [15] performed the best, with a fidelity of 25%. The substantial error of these previous models is a likely reflection of both missing parameters and the challenge of finding the exact functional form required for predicting the pupil diameter. Later models added variables such as adaptation field, age, and monocular adaptation [2,16-21]. All of the earlier models considered ambient light levels by way of the total luminance as opposed to the fine wavelength resolution of the UV/visible spectrum that was used in this study. The fine wavelength resolution allows one to identify the wavelengths to which the pupil dilation is most sensitive, it is noteworthy that there are some small variations from eye to eye in the key wavelengths for determining the pupil diameter. In this study we have utilized recent technological developments, the full visible spectrum and pupil size can be measured with high accuracy and in large volume combined with machine learning, this provides new opportunities for the development of much more robust higher fidelity empirical models.

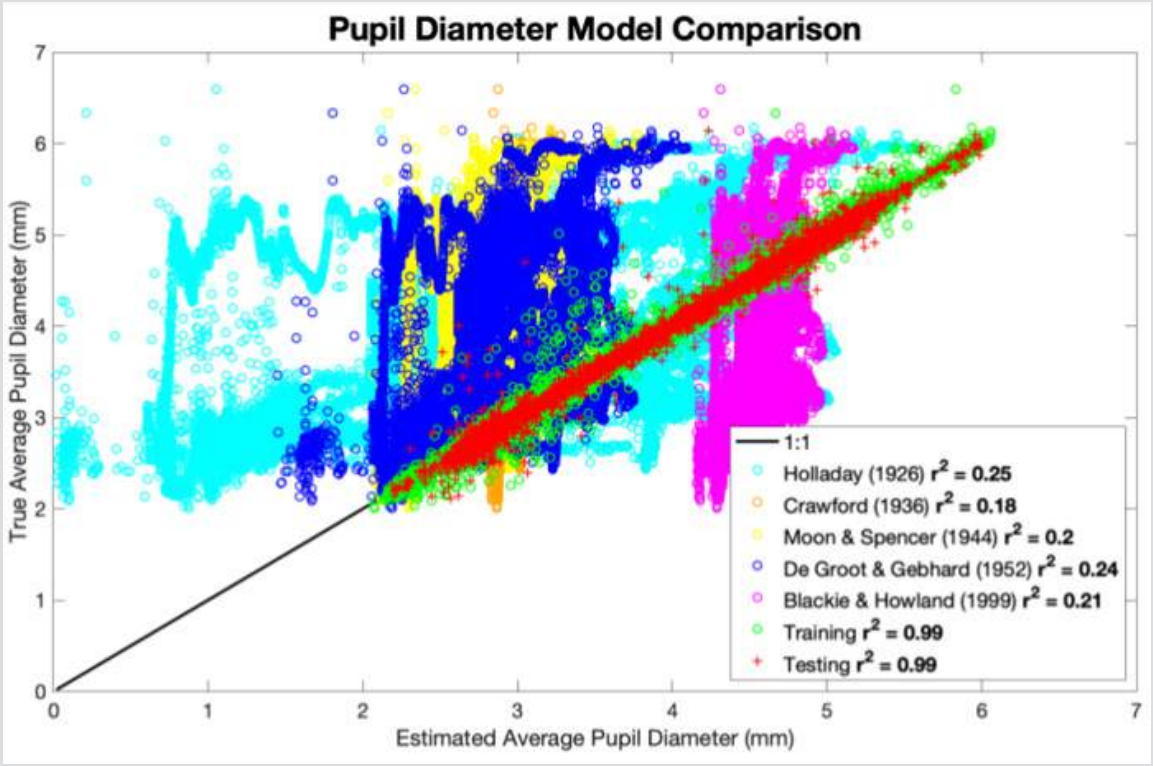


Figure 1: Evaluation and comparison of previous pupil diameter models which utilized a single variable, luminance, showing poor fidelity contrasted with the multivariate empirical machine learning model for the average pupil diameter developed in this study showing good fidelity (foreground green training and red validation points). The true average diameter of the left and right pupils is given on the y-axis, and the estimation by each respective model on the x-axis. Luminance was computed from measured illuminance where the luminance was assumed to be isotropic and reflectance assumed to be 1. Models were evaluated based on description by Watson and Yellott [2].

In this first demonstration case study, with just one participant, we examined the effect of both light intensity and the orientation/motion of the head on the diameter of a participant's pupils. Different illumination environments can be characterized by their spectra. This light consisting of various wavelengths can interact with different photo-receptors (light sensitive cones) in the retina. This interaction produces electrical signals that are sent to the brain and interpreted as color [22]. These cones are disproportionately sensitive to particular wavelengths with absorbance peaks around 420 nm (violet), 534 nm (green), and 564 nm (yellow-green) [3]. An illustration of these sensitivities can be shown by a plot of the mean absorbance of the three classes of photo-receptors (short-wave, middle-wave, and long-wave cones) vs wavelength (Figure 2).

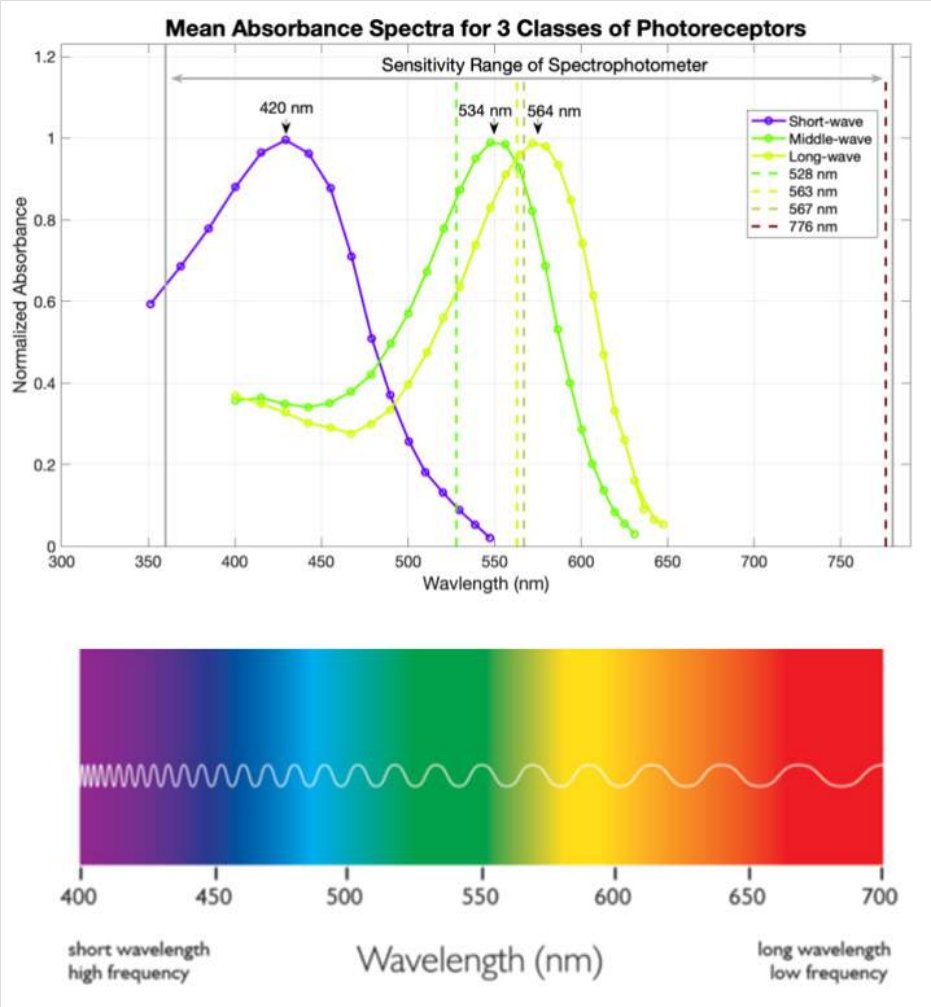


Figure 2: Normalized mean absorbance spectra for long-wave, middle-wave, and short-wave cones. Maximum absorbance values for each class of cones are 420 nm ±4.7 nm, 534 nm ±3.7 nm, and 564 ±4.7 nm, 420 nm ±4.7 nm, respectively. Dashed vertical lines represent the top 4 important predictors taken from the pupil diameter models created here. The sensitivity range of the Konica Minolta CL- 500A Spectrophotometer is 360 – 780 nm indicated by the gray double-sided arrow. Cone absorbances were based on a figure in the paper by Bowmaker and Dartnall [3].

New predictive empirical models of the pupil diameter can be derived using supervised multivariate non-linear non-parametric machine learning regression. The accuracy of the models can be evaluated using an independent validation (or testing) dataset whose data records were not utilized in the model training. This machine learning approach can also provide insights on the relative importance of the inputs (i.e. predictors). In this case we had a few hundred inputs, including the light intensities for every nm of wavelengths from 360-780 nm (ultra-violet to near infrared).

Materials and Methods

Data was collected during 3 outdoor/indoor walks where spectral and biometric data were recorded. The walks took place in the morning (8:30 AM) and late afternoons (4 PM), each lasting approximately fifteen minutes. Spectral data was measured approximately every 3 seconds using a NIST calibrated Konica Minolta CL-500A Illuminance Spectrophotometer, which measures the illuminance and spectral irradiance of wavelengths from 360-780 nm with 1±0.3 nm resolution. Pupil diameters, head orientation, and the proper acceleration of the head were recorded 100 times a second using Tobii Pro Glasses 2. The glasses use an infrared grid projected onto each eye to estimate the position and size of the pupils. The orientation and acceleration of the head are estimated using a Microelectromechanical System (MEMS) gyroscope and MEMS accelerometer located in the glasses. Data was prepared and analyzed using Matlab 2019a.

The data preparation involved six steps:

1. **Collection** - Recording of the raw data. Data was written to 6 separate files corresponding to the 2 devices for each of the 3 trials.

2. **Formatting** - Converting raw data files to Matlab timetable objects. 6 timetables were created from the raw data files.

3. Synchronizing - The sampling frequencies differed for each device. 1 record every 3 seconds for the spectral data, versus 100 records every second for the biometric data. To account for this, the 2 timetables for a particular trial were reconfigured to share the same time steps using Matlab's retime function with a linear interpolation. The timetables for each trial could then combined using the synchronize function. Resulting in 3 timetables, one for each of the 3 trials.

4. Merging - Concatenating all 3 timetables into a single timetable.

5. Cleaning - Removing records with device error flags, NaN elements, and zero values for pupil diameter. The latter case is addressed below.

6. Generating - Creating new variables such as the average pupil diameter and inter-eye pupil diameter difference.

A major challenge was introduced in step 5 (cleaning) of the data preparation due to a significant portion of the pupil diameter records taking values of 0. This was a non-physical consequence of the mechanism with which the pupil diameters were measured. When there is a high intensity of ambient infrared light from bright sunshine the glasses can no longer readily discern the pupil diameter, this is reflected in Figure 3 where pupil diameter dropouts coincide with time intervals of high spectral irradiance. These records were removed from the data, reducing the number of records from 380,000 to 80,000 records.

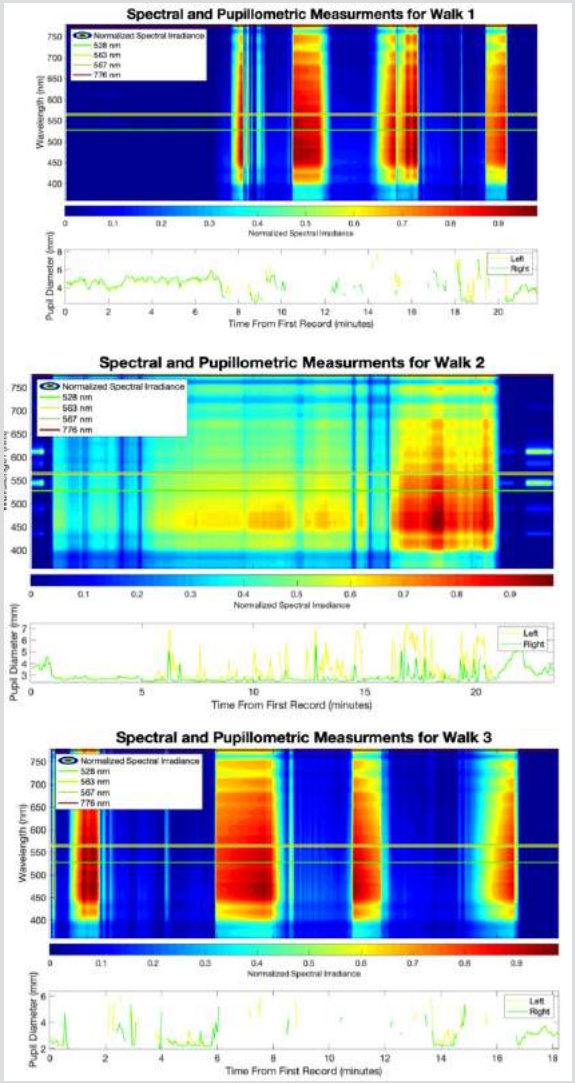


Figure 3: The normalized spectral irradiance at every time step for all walks is plotted. The irradiance is normalized by dividing all values by the maximum spectral irradiance within each walk. Relative size of irradiance values are indicated by the colorbar. Spectral lines at 528, 563, 567, and 776 nm represent the most important predictors for the pupil diameter models. Left (yellow) and right (green) pupil diameters are plotted over time. Note the pupil diameter dropouts in time intervals where the spectral irradiance is high.

a) Walk 1 measurements during late afternoon (≈ 4PM).

b) Walk 2 measurements during morning (≈ 8:30 AM) with overcast.

c) Walk 3 measurements during late afternoon (≈ 4 PM).

From the recorded data we sought to estimate 5 different parameters, namely the: average of the left and right pupil diameters (APD), left pupil diameter (LPD), right pupil diameter(s) (RPD), magnitude of the difference between the left and right pupil diameters (PDD), and the illuminance. These parameters can be estimated by constructing objectively optimized empirical machine learning models. The hyperparameters (i.e. the parameters that define options associated with the training process) of an ensemble of regression trees able to use both boosting and bagging were optimized (the Matlab function fitrensemble with the Optimize Hyperparameters option set to all). More information on this function is available in the Matlab documentation [23]. We have done many previous machine learning studies [24-56]. The data was split into 2 subsets: one for training and one for the independent testing of each empirical machine learning model. With 90% of the data used for training the multivariate non-linear non-parametric regression models and 10% of the data used for independent testing of the models.

Results and Discussion

In the following subsections we discuss the results of the 5 different empirical machine learning models. The accuracy of each model was assessed via a scatter plot of the true vs estimated response variable values (see Figures 4a, 5a, 6a, 7a, & 9a). If the true and estimated values are identical, the resulting scatter plot will be a straight line with a slope of one and an intercept of zero, i.e. a perfect one to one plot with a correlation coefficient, r^2, equal to 1. This ideal is indicated by a black line in each scatter plot. The correlation coefficients for the training (plotted as green circles) and testing (plotted as red pluses) datasets were computed using Matlab's corrcoef function.

The relative predictor importance ranking of each model was derived using the predictorImportance function. The relative rankings are visualized as bar plots (see Figures 4b, 5b, 6b, 7b, & 9b). The importance estimates are plotted on a log scale with the most important predictors shown toward the top. In the pupil diameter models (i.e. models for the APD, LPD, RPD, and PDD), the top 20 out of 427 predictors are shown. For the illuminance model, all 7 predictors are given in the ranking. The top 3 predictors are indicated by red bars, the next 2 important predictors by yellow bars, and the remaining predictors by blue bars.

The Average Pupil Diameter Model

Figure 4 shows the results of the Average Pupil Diameter (APD) model. The APD was estimated using the spectral irradiance at every nm between 360-780 nm, the gyroscope, and the accelerometer data as predictor variables. The scatter plot of the true vs the estimated average pupil diameter values is shown in Figure 4a. The model had correlation coefficients of > 0.99 for both the training and testing data subsets. Thus, the empirical machine learning model was successful in predicting the average pupil diameter. Figure 4b shows the ranking of the relative importance of the inputs in predicting the APD, the top 3 predictors are the irradiance values at 561, 563, and 562 nm, which coincides with the maximum absorbance of the long-wave cones at around 563 nm [3]. This suggests the long-wave photo-receptors play a more significant role than the short- or middle-wave receptors in controlling the average size of the pupils for the participant.

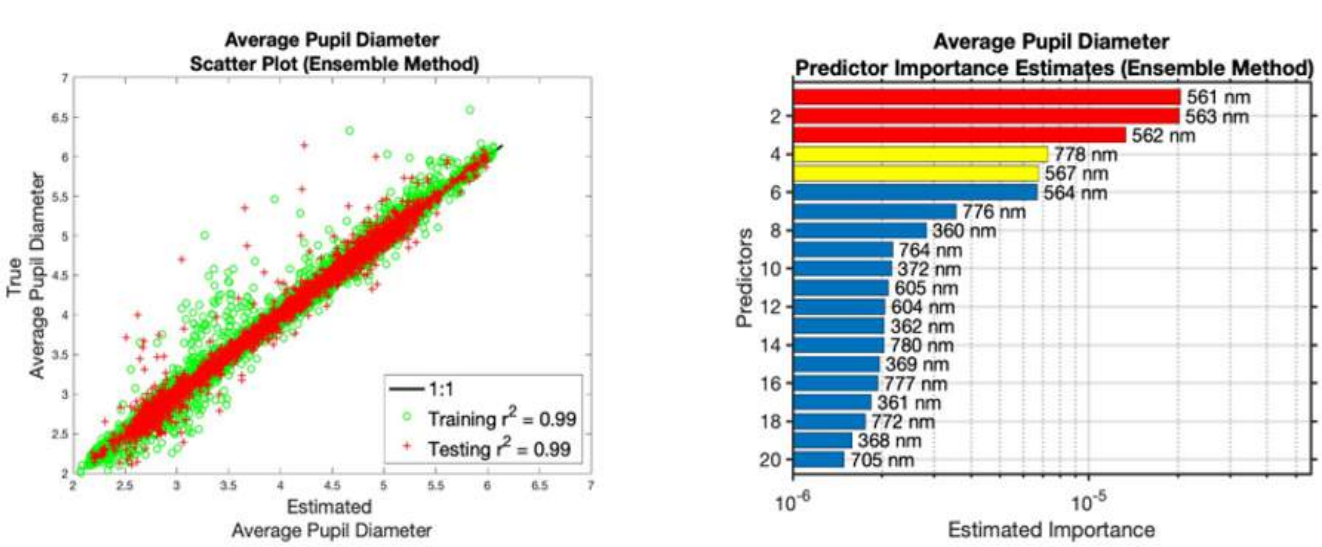


Figure 4: Plots for the Average Pupil Diameter model.

a) True vs estimated average pupil diameter in millimeters.

b) Predictor importance estimates for the average pupil diameter model.

The Left Pupil Diameter Model

The results for the Left Pupil Diameter (LPD) model are shown in Figure 5. The LPD was estimated using the same predictors as the APD, the spectral irradiance from 360-780 nm, the gyroscope, and the accelerometer data. The model was successful in predicting the LPD with a correlation coefficient of > 0.96 for both the training and validation data subsets. The top predictor (567 nm) is again near the maximum absorbance of the long-wave photo-receptors (563 nm). The next top 6 predictors are the irradiance values at 528, 568, 564, 527, 668 and 570 nm, which seem to coincide with both the middle- and long-wave photo-receptors with maximum absorbance values near 533.8 ± 3.7 nm and 563 nm, respectively, with the exception of the irradiance at 668 nm [3].

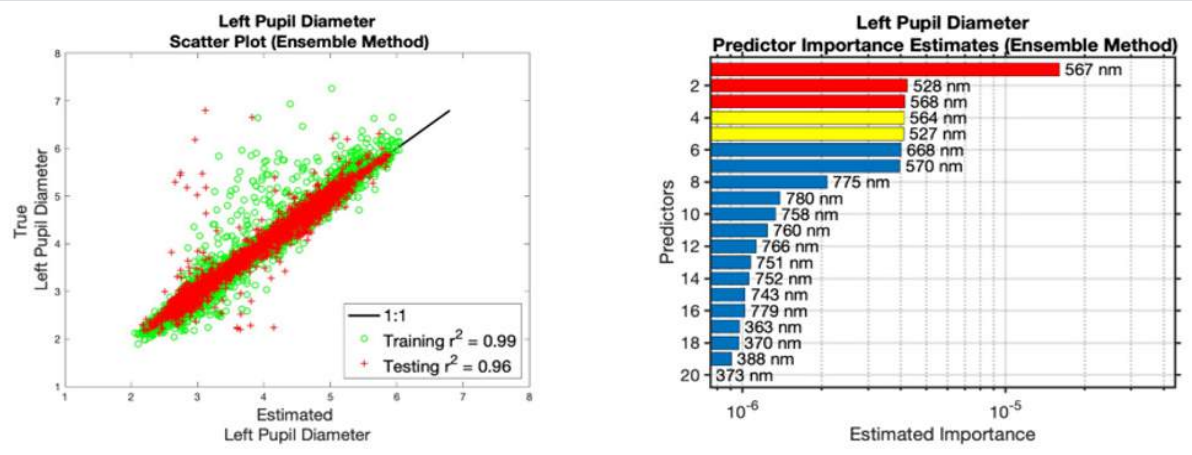


Figure 5: Plots for the Left Pupil Diameter model.

a) True vs estimated left pupil diameter in millimeters.

b) Predictor importance estimates for the left pupil diameter model.

The Right Pupil Diameter Model

The results for the Right Pupil Diameter (RPD) model are shown in Figure 6. The RPD was estimated using the same predictors as the APD and LPD. For the RPD model there is a strong correlation between the estimated and true values, with coefficients of determination > 0.99 for both data subsets, shown in Figure 6a. The top 2 predictors are 563 nm and 562 nm, which again coincide with the maximum absorbance of the long-wave cones near 563 nm. The next most important predictor was the irradiance at 776 nm corresponding to near infrared light. This and the appearance of near infrared predictors in all the importance rankings may be a consequence of the infrared noise in the environment, resulting in the measured pupil diameters to be smaller than the actual values. An interesting result from the importance ranking in Figure 6b, is the appearance of a non-spectral predictor (Accelerometer Z) which denotes the proper acceleration in the direction in front of the glasses. This may be correlated to the participant looking down to navigate obstacles in the walking path such as stairs, inclines, rugged terrain, and other impediments. Focusing on a specific task or object may cause an increase in cognitive load, resulting in a pupillary response [10,11].

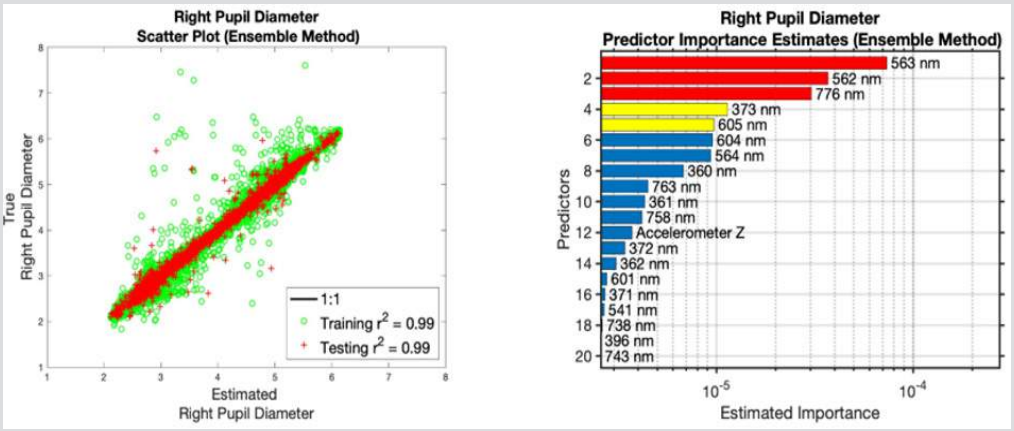


Figure 6: Plots for the Right Pupil Diameter model in millimeters.

a) True vs. estimated right pupil diameter in millimeters.

b) Predictor importance estimates for the right pupil diameter model.

The Pupil Diameter Difference Model and Pupil Asymmetry

The results for the left and right pupil diameter models are noticeably different (see Figures 5 and 6), which may suggest an asymmetry in the behavior of each pupil. One measure of this asymmetry is the magnitude of the difference between the left and right pupil diameters. This is shown by the results of the Pupil Diameter Difference (PDD) model given in Figure 7. The same predictors were used for the PDD model as in the APD, LPD, and RPD models. This empirical model was not successful in predicting the PDD, since the correlation coefficient was 0.43 for the testing data subset, as shown in Figure 7a. Clearly the most important predictors for modeling this asymmetry were not available in the training dataset.

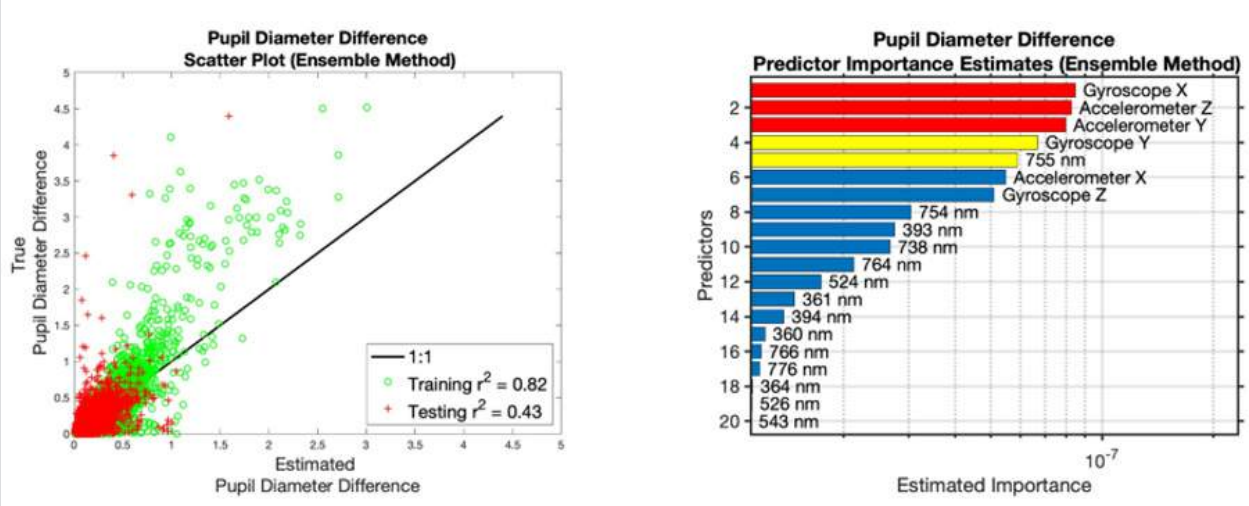


Figure 7: Plots for the Pupil Diameter Difference .
a) True vs estimated pupil diameter differences in millimeters.
b) Predictor importance estimates for the pupil diameter difference model.

Another metric of the pupil asymmetry can be the accuracy of the LPD model in estimating the RPD and vice versa. The resulting scatter plots are given in Figure 8. Despite the differences in the importance rankings and failures of the PDD model, the estimates are fairly accurate with correlation coefficients of > 0.95 for both the testing and training datasets. This accuracy may suggest that although there is an asymmetry in the importance rankings for the left and right pupil models, the functioning of each pupil is very similar. A possible cause of this asymmetry is ocular dominance (i.e. the input for one eye is preferred over the other) [57,58]. It has been suggested that ocular dominance is not a static phenomenon, but will vary with changing horizontal gaze angle [59].

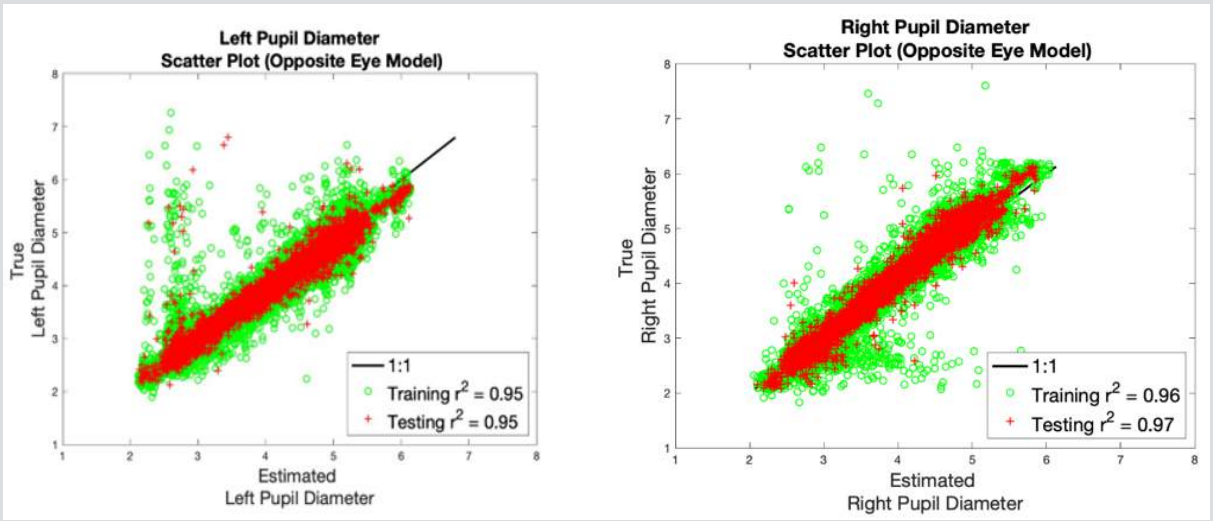


Figure 8: Plots for the pupil diameter prediction using model from opposite eye data. Pupil diameters are in millimeters.
a) True vs estimated left pupil diameter using the right pupil diameter model.
b) True vs estimated right pupil diameter using the left pupil diameter model.

The Illuminance Model

Figure 9 shows the results of the Illuminance model. We just saw above that if we know the light intensity we can accurately predict the pupil diameter, so now we `invert' the experiment and ask the question, if we know the pupil diameter can we accurately estimate the light intensity? The model used the pupil diameters, gyroscope, and accelerometer data as the predictors. The estimates were somewhat accurate with correlation coefficients of 0.91 and 0.71 for the training and testing datasets, respectively. The top 2 predictors are the left and right pupil diameters, which agrees with first order considerations of the relationship between pupil diameters and external light levels. The next most important predictor was the acceleration in the z-direction (forward direction). Which may again be correlated with participant focus on obstacle navigation.

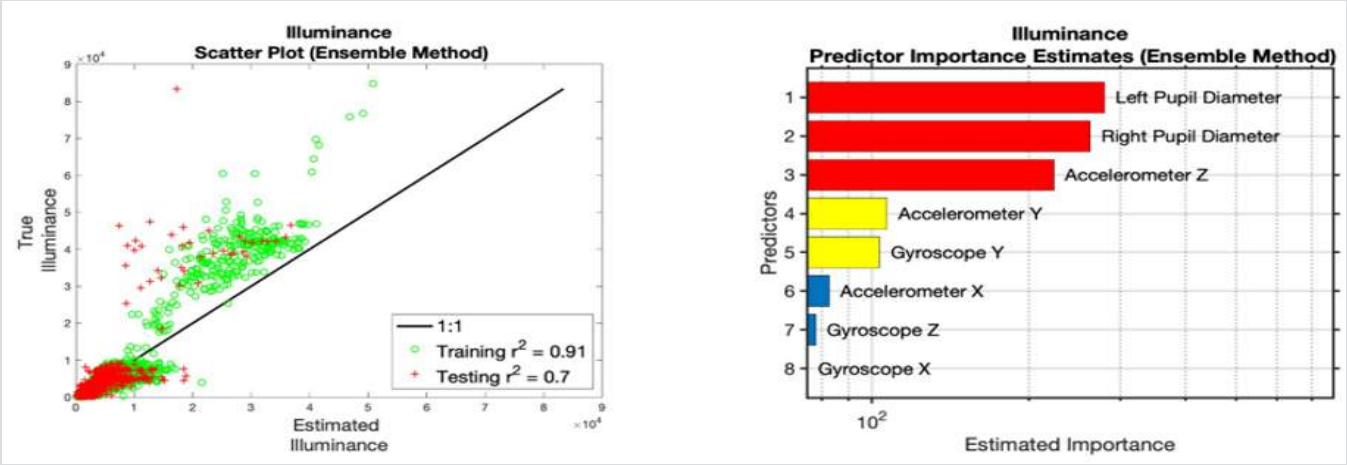


Figure 9: Plots for the Illuminance model.
a) True vs estimated illuminance in lux.
b) Predictor importance estimates for the illuminance model.

Pupil Diameter and Illuminance

In a first order consideration, we can expect the pupil diameter to be inversely proportional to the illuminance. This is depicted in Figure 10, which gives 3 scatter plots of the average, left, and right pupil diameters vs illuminance. At low illuminance values, the expected inverse relationship is apparent. At higher values (> 4000 lux) this expectation fails. The lack of a clear relationship between the two variables in all situations is likely the main contributor to the failure of previous models (Figure 1).

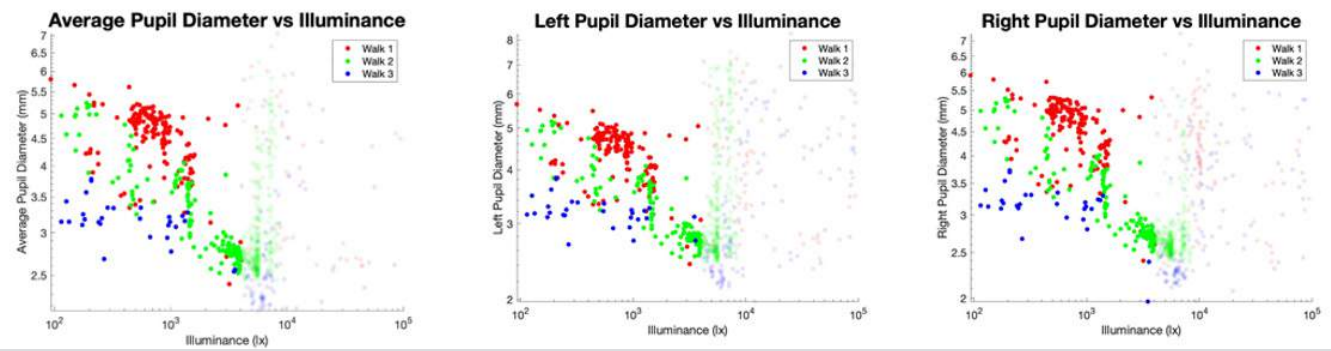


Figure 10: Log scale scatter plots of the pupil diameters vs illuminance. Data from walks 1, 2, and 3 are distinguished by the colors red, green, and blue, respectively. Data points with low opacity have illuminance values above 4000 lx. Note below the 4000 lx mark the variables tend to have an inverse relationship.
a) Average pupil diameter vs illuminance.
b) Left pupil diameter vs illuminance.
c) Right pupil diameter vs illuminance.

The Environment

The normalized spectral irradiance at every time step for each trial is given in Figure 3. Normalized values were computed by dividing all irradiance values by the largest irradiance within each trial. Spectral lines are plotted for 528, 563, 567, and 776 nm, based on the top 3 most important predictors across all pupil diameter models (see Figures 4b, 5b, 6b, & 7b). Where predictors of the spectral irradiance at 561, 562, and 568 nm were disregarded in lieu of the irradiance at 563 and 567 nm.

Temporal discontinuities in the spectra are due to those time intervals in which the participant walked in and out of shaded areas and/or away from the sun, which resulted in orders of magnitude differences in the spectral irradiance. Figure 11 depicts the normalized spectral irradiance plotted on a log scale. Time intervals colored predominately red represent outdoor spectra, while more colorful intervals are indoor.

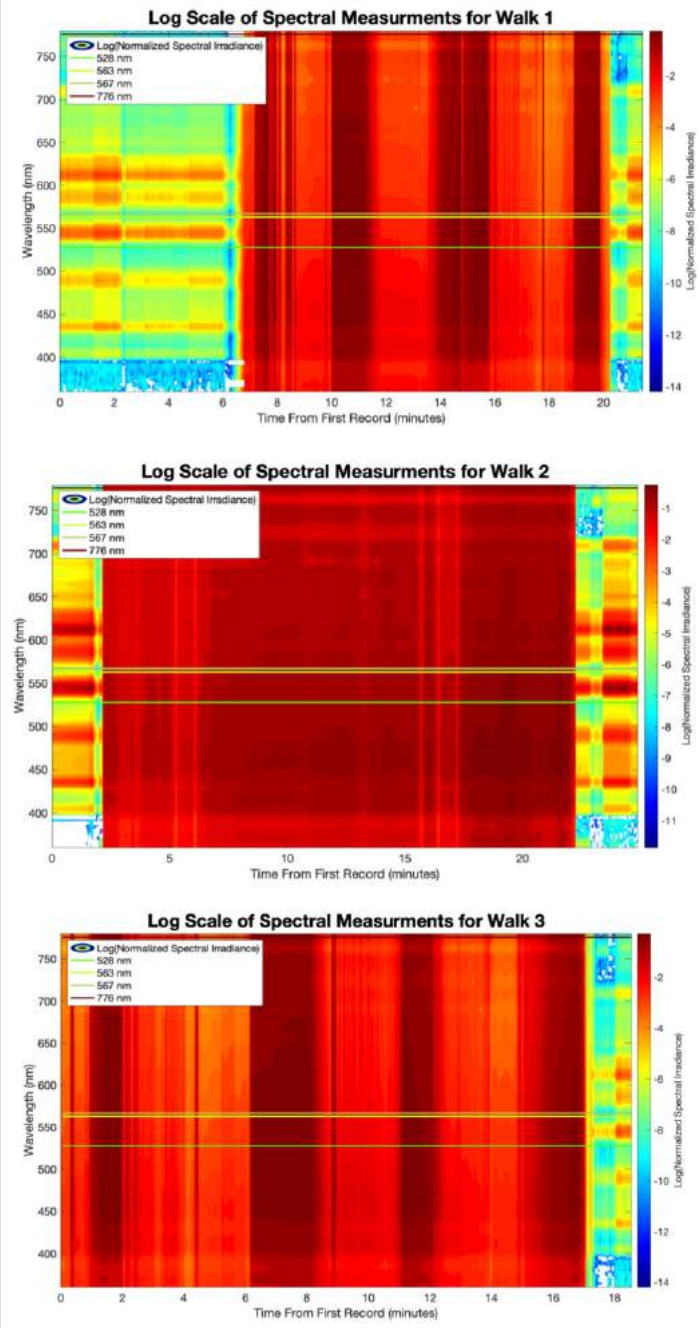


Figure 11: The log of the normalized spectral irradiance at every time step for all walks is plotted. The irradiance is normalized prior to taking log by dividing all values by the maximum spectral irradiance within each walk. Relative sizes of irradiance values are indicated by the color bar. Spectral lines at 528, 563, 13, 567, and 776 nm represent wavelengths of the most important predictors for the pupil diameter models.

a) Walk 1 measurements during late afternoon (≈ 4PM).

b) Walk 2 measurements during morning (≈ 8:30 AM).

c) Walk 3 measurements during late afternoon (≈ 4PM).

Limitations

The high level of infrared noise caused significant drawbacks in the data analysis. Further developments may require light intensities and spectra to be within a non-disruptive range. Another solution may be to utilize an eye tracking instrument which uses visible light to estimate the pupil diameters.

Future Directions

Pupil size along with other autonomic responses such as heart rate variability, galvanic skin response, and core temperature changes have been associated with cognitive load and performance [5-11]. Although cognitive load is a significant contributor to the provocation of these responses, in a dynamic outdoor environment and while performing a physical activity (such as walking or cycling) it is not always clear which responses were due to external stimuli or cognitive status. Using a similar approach to the one used here, future data collection will expand the number of participants, environments, cognitive tasks, and biometric sensors.

Looking forward, multiple participants will allow for the assessment of the inter-person variability of the models, including parameters such as age and body composition. Different environments will vary in light intensity, air quality, elevation, and temperature. Environmental variables can be measured using mobile weather stations mounted on a participant or bicycle. Other environmental sensors such as a video camera, microphone, and LIDAR can indicate dynamic field situations and track events. Tasks such as walking and cycling will be performed. Cyclist performance can be assessed via bicycle speed and biometric data. Biometrics such as Electroencephalography (EEG), Heart Rate (ECG), Galvanic Skin Response (GSR), body temperature, Electromyography (EMG), blood oxygen level, and respiration will be considered and modeled. The ranking of predictor importance for these biometric models can help identify important relationships between environmental stimuli and different autonomic response.

Conclusion

Past formulae for predicting pupil diameter mainly considered total ambient light levels via luminance [2,15-21], these models could not capture the fully multi-variate and non-linear dependence of pupil diameter on the environmental state, and consequently had poor generalization. When considering the spectrum of light from 360-780 nm (ultra-violet to near infrared) in lieu of the luminance, we were able to derive a very accurate empirical machine learning model which can predict pupil diameters with a minimum fidelity of 96.9%. The machine learning also allowed us to identify that the most important wavelengths in predicting the pupil diameters were around 562 nm (green), which is near the peak absorbance of the long-wave photo-receptive cones (562.8 ± 4.7 nm) [3].

Funding

This research was funded by USAMRMC Award Number W81XWH-18-1-0400.

References

1. Reeves P (1920) The response of the average pupil to various intensities of light. J Opt Soc Am 4: 35-43.
2. Watson AB, Yellott JI (2012) A unified formula for light-adapted pupil size. Journal of vision 12: 12.
3. Bowmaker JK, Dartnall HJ (1980) Visual pigments of rods and cones in a human retina. J Physiol 298: 501-511.
4. Laeng B, Sirois S, Gredeback G (2012) Pupillometry: A window to the preconscious? Perspectives Psychological Science 7: 18-27.
5. Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH (2018) Stress and heart rate variability: A meta-analysis and review of the literature. Psychiatry Investig 15: 235-245.
6. Vetrugno R, Liguori R, Cortelli P, Montagna P (2003) Sympathetic skin response. Clinical Autonomic Research 13: 256-270.
7. Taylor L, Watkins SL, Marshall H, Dascombe BJ, Foster J (2015) The impact of different environmental conditions on cognitive function: A focused review. Frontiers Physiology 6: 372.
8. Kahneman D, Beatty J (1966) Pupil diameter and load on memory. Science 154: 1583-1585.
9. Hess EH, Polt JM (1964) Pupil size in relation to mental activity during simple problem-solving. Science 143: 1190-1192.
10. Wel VD P, Steenbergen VH (2018) Pupil dilation as an index of e ort in cognitive control tasks: A review. Psychonomic Bulletin Review 25: 2005-2015.
11. Mathot S (2018) Pupillometry: Psychology, physiology, and function. Journal of Cognition 1(1): 16.
12. Winn B, Whitaker D, Elliott DB, Phillips NJ (1994) Factors affecting light-adapted pupil size in normal human subjects. Investigative ophthalmology visual science 35: 1132-1137.
13. Richard CC, Oka S, Bradshaw DH, Jacobson RC, Donaldson GW(1999) Phasic pupil dilation response to noxious stimulation in normal volunteers: Relationship to brain evoked potentials and pain report. Psychophysiology 36: 44-52.
14. Hess EH, Polt JM (1960) Pupil size as related to interest value of visual stimuli. Science 132: 349-350.
15. Holladay LL (1926) The fundamentals of glare and visibility. J Opt Soc Am 12: 271-319.
16. Crawford BH (1936) The dependence of pupil size upon external light stimulus under static and variable conditions. Proceedings of the Royal Society B-Biological Sciences 121: 376-395.
17. Moon P, Spencer DE. On the stiles-crawford effect. J Opt Soc Am 34: 319-329.
18. de Groot SG, Gebhard JW (1952) Pupil size as determined by adapting luminance. J Opt Soc Am 42: 492-495.
19. Blackie CA, Howland HC (1999) An extension of an accommodation and convergence model of emmetropization to include the effects of illumination intensity. Ophthalmic physiological optics 19: 112-125.
20. Stanley PA, Davies AK (1995) The effect of field of view size on steady-state pupil diameter. Ophthalmic Physiological Optics 15: 601-603.
21. Peter GJ Barten (1999) Contrast sensitivity of the human eye and its effect on image quality.
22. Williamson SJ, Cummins HZ (1983) Light and color in nature and art. Wiley 10: 123-124.
23. Works M Matlab Documentation kernel description.
24. Lary DJ, Alavi AH, Gandomi AH, Walker AL (2016) Machine learning in geosciences and remote sensing. Geoscience Frontiers 7: 3-10.
25. Brown ME, Lary DJ, Vrieling A, Stathakis D, Mussa H (2018) Neural networks as a tool for constructing continuous NDVI time series from AVHRR and MODIS. International Journal of Remote Sensing 29: 7141-7158.

26. Lary DJ, Remer LA, MacNeil D, Roscoe B, Paradise S (2009) Machine learning and bias correction of MODIS aerosol optical depth. IEEE Geoscience and Remote Sensing Letters 6: 694-698.
27. Lary DJ, Aulov O (2008) Space-based measurements of HCL: Intercomparison and historical context. Journal of Geophysical Research: Atmospheres 113.
28. Lary DJ, Muller MD, Mussa HY (2004) Using neural networks to describe tracer correlations. Atmospheric Chemistry and Physics 4: 143-146.
29. Malakar NK, Lary DJ, Gencaga D, Albayrak A, and Wei J (2013) Towards identification of relevant variables in the observed aerosol optical depth bias between MODIS and AERONET observations. In AIP Conference Proceedings 1553: 69-76.
30. David John Lary (2010) Artificial intelligence in aerospace. In aerospace technologies advancements.
31. Malakar NK, Lary DJ, Moore A, Gencaga D, Roscoe B, et al. (2012) Estimation and bias correction of aerosol abundance using data-driven machine learning and remote sensing. In 2012 Conference on Intelligent Data Understanding 24-30.
32. Lary DJ (2013) Using multiple big datasets and machine learning to produce a new global particulate dataset: A technology challenge case study. In AGU Fall Meeting Abstracts.
33. Lary D (2007) Using neural networks for instrument cross-calibration. In AGU Fall Meeting Abstracts.
34. Albayrak A, Wei JC, Petrenko M, Lary DJ, Leptoukh GG (2011) Modis aerosol optical depth bias adjustment using machine learning algorithms. In AGU Fall Meeting Abstracts.
35. Brown ME, Lary DJ, Mussa H (2006) Using neural nets to derive sensor-independent climate quality vegetation data based on AVHRR, SPOT-vegetation, seaWiFS and MODIS. In AGU Spring Meeting Abstracts.
36. Lary DJ, Muller MD, Mussa HY (2003) Using neural networks to describe tracer correlations. Atmospheric Chemistry and Physics Discussions 3: 5711-5724.
37. Malakar NK, Lary DJ, Allee R, Gould R, Ko D (2012) Towards automated ecosystem-based management: A case study of northern gulf of mexico water. In AGU Fall Meeting Abstracts.
38. Lary DJ (2014) Bigdata and machine learning for public health. In 142nd APHA Annual Meeting and Exposition 2014.
39. Lary DJ, Lary T, Sattler B (2015) Using machine learning to estimate global $pm_{2.5}$ for environmental health studies. Environmental Health Insights 12: 41-52.
40. Kneen MA, Lary DJ, Harrison WA, Annegarn HJ, Brikowski TH. (2016) Interpretation of satellite retrievals of $pm_{2.5}$ over the southern african interior. Atmospheric Environment 128: 53-64.
41. Lary DJ, Nikitkov A, Stone D, (2010) Which machine-learning models best predict online auction seller deception risk. American Accounting Association AAA Strategic and Emerging Technologies.
42. Medvedev IR, Schueler R, Thomas J, Kenneth O, Nam HJ, et al. (2016) Analysis of exhaled human breath via terahertz molecular spectroscopy. In 2016 41st International Conference on Infrared, Millimeter, and Terahertz waves (IRMMW-THz) 1-2.
43. Lary DJ, Lary T, Sattler B (2016) Using machine learning to estimate global $PM_{2.5}$ for environmental health studies. Geoinformatics & Geostatistics: An Overview 4(4).
44. KK O, Zhong Q, Sharma N, Choi W, Schueler R, et al. (2017) Demonstration of breath analyses using cmos integrated circuits for rotational spectroscopy. In International Workshop on Nanodevice Technologies, Hiroshima, Japan.
45. Wu D, Zewdie GK, Liu X, Kneed M, Lary DJ (2017) Insights into the morphology of the east asia $pm_{2.5}$ annual cycle provided by machine learning. Environmental Health Insights 11: 1-7.
46. Nathan BJ, Lary DJ (2019) Combining domain filling with a self-organizing map to analyze multi-species hydrocarbon signatures on a regional scale. Environmental Modeling and Assessment 191: 337.
47. Lary MA, Allsop L, Lary DJ (2019) Using machine learning to examine the relationship between asthma and absenteeism. Environmental Modeling and Assessment 191: 332.
48. Lary DJ, Zewdie GK, Liu X, Wu D, Levetin E, et al. (2018) Machine learning applications for earth observation. In Earth Observation Open Science and Innovation. ISSI Scienti c Report Series 15: 165-218.
49. Wu D, Lary DJ, Zewdie GK, Liu X (2019) Using machine learning to understand the temporal morphology of the $pm_{2.5}$ annual cycle in east asia. Environmental Monitoring and Assessment 191: 272.
50. Alavi AH, Gandomi AH, Lary DJ. Progress of machine learning in geosciences.
51. Ahmad Z, Choi W, Sharma N, Zhang J, Zhong Q, et al. (2016) Devices and circuits in CMOS for THz applications. In 2016 IEEE International Electron Devices Meeting (IEDM) 29: 8.
52. Zewdie G, Lary DJ (2018) Applying machine learning to estimate allergic pollen using environmental, land surface and NEXRAD radar parameters. In AGU Fall Meeting Abstracts.
53. Malakar NK, Lary DJ, Gross B (2018) Case studies of applying machine learning to physical observation. In AGU Fall Meeting Abstracts.
54. Zewdie GK, Lary DJ, Levetin E, Garuma GF (2019) Applying deep neural networks and ensemble machine learning methods to forecast airborne ambrosia pollen. International journal of environmental research and public health 16(11): E1992.
55. Zewdie GK, Lary DJ, Liu X, Wu D, Levetin E (2019) Estimating the daily pollen concentration in the atmosphere using machine learning and NEXRAD weather radar data. Environmental Monitoring and Assessment 191(7): 418.
56. Chang HH, Pan A, Lary DJ, Waller LA, Zhang L (2019) Time-series analysis of satellite-derived ne particulate matter pollution and asthma morbidity in jackson, MS. Environmental Monitoring and Assessment 191: 280.
57. Miles WR (1930) Ocular dominance in human adults. The Journal of General Psychology 3(3): 412-430.
58. Porac C, Coren S (1976) The dominant eye. Psychological bulletin 83(5): 880-897.
59. Khan AZ, Crawford JD (2001) Ocular dominance reverses as a function of horizontal gaze angle. Vision Research 41(14): 1743-1748.

ISSN: 2574-1241

DOI: 10.26717/BJSTR.2019.20.003446

Shawhin Talebi. Biomed J Sci & Tech Res

sensors

MDPI

Article

Using Machine Learning for the Calibration of Airborne Particulate Sensors

Lakitha O.H. Wijeratne *, Daniel R. Kiv, Adam R. Aker, Shawhin Talebi and David J. Lary

University of Texas at Dallas, 800 W, Campbell Rd, Richardson, TX 75080, USA; drk150030@utdallas.edu (D.R.K.); Adam.Aker@utdallas.edu (A.R.A.); Shawhin.Talebi@utdallas.edu (S.T.); David.Lary@utdallas.edu (D.J.L.)

* Correspondence: lhw150030@utdallas.edu

Received: 7 November 2019; Accepted: 10 December 2019; Published: 23 December 2019

Abstract: Airborne particulates are of particular significance for their human health impacts and their roles in both atmospheric radiative transfer and atmospheric chemistry. Observations of airborne particulates are typically made by environmental agencies using rather expensive instruments. Due to the expense of the instruments usually used by environment agencies, the number of sensors that can be deployed is limited. In this study we show that machine learning can be used to effectively calibrate lower cost optical particle counters. For this calibration it is critical that measurements of the atmospheric pressure, humidity, and temperature are also made.

Keywords: optical particle counter; airborne particulates; machine learning

1. Introduction

Airborne atmospheric aerosols are an assortment of solid or liquid particles suspended in air [1]. Aerosols, also referred to as particulate matter (PM), are associated with a suite of issues relevant to the global environment [2–8], atmospheric photolysis, and a range of adverse health effects [9–15]. Atmospheric aerosols are usually formed either by direct emission from a specific source (e.g., combustion) or from gaseous precursors [16]. Although individual aerosols are typically invisible to the naked eye, due to their small size, their presence in the atmosphere in substantial quantities means that their presence is usually visible as fog, mist, haze, smoke, dust plumes, etc. [17]. Airborne aerosols vary in size, composition, origin, and spatial and temporal distributions [14,18]. As a result, the study of atmospheric aerosols has numerous challenges.

1.1. Motivation for This Study

Low cost sensors that can also be accurately calibrated are of particular value. For the last two decades we have pioneered the use of machine learning to cross-calibrate sensors of all kinds. This was initially done for very expensive orbital instruments onboard satellites (awarded an IEEE paper prize, and specially commended by the NASA MODIS team) [19]. We are now using this approach operationally for low-cost sensors distributed at scale across dense urban environments as part of our smart city sentinels. The approach can be used for very diverse sensors, but as a useful illustrative example that has operational utility, we describe here a case for accurately calibrated, low-cost sensors measuring the abundance and size distribution of airborne particulates, with the implicit understanding that many other sensor types could easily be substituted. These sensors can be readily deployed at scale at fixed locations; be mobile on various robotic platforms (walking, flying, etc) or vehicles; be carried; or deployed autonomously as a mesh network, either by operatives or by robots (walking, flying, etc.).

Building in calibration will enable consistent data to retrieved from all the low-cost nodes deployed/thrown. Otherwise the data will always be under some suspicion as the inter-sensor

variability among low-cost nodes can be substantial. While much effort has been recently placed on providing the connectivity of large disbursed low-cost networks, little to no effort has been spent on the automated calibration, bias-detection, and uncertainty estimation necessary to make sure the information collected is sound. A case study of providing this critical calibration using machine learning is the focus of this paper.

Any sensor system benefits from calibration, but low-cost sensors are typically in particular need of calibration. The inter-sensor variability among low-cost nodes can be substantial. In addition, to the pre-deployment calibration, once the sensors have been deployed, the paradigm we first developed for satellite validation of constructing probability distribution functions of each sensor's observation streams, can be used to both monitor the real-time calibration of each sensor in the network by comparing its readings to those of its neighbors, but also to answer the question "how representative is an instantaneous reading of the conditions seen over some temporal and spatial window within which the sensor is placed?".

1.2. Using Probability Distribution Functions to Monitor Calibration and Representativeness in Real-Time

It is useful to be able to answer the question, "How representative is an instantaneous reading of the conditions seen over some temporal and spatial window within which the sensor is placed?". We can answer this question by considering a probability distribution functions (PDFs) of all the observations made by a sensor over some temporal and spatial window [20]. The width of this probability distribution is termed the representativeness uncertainty for that temporal and spatial window. The PDFs of all observations made by each sensor are automatically compared in real time to the PDFs from the neighboring sensors within a neighborhood radius. These neighborhood sensors can include measurements from primary reference sensors that may be available. This comparison is used to estimate the measurement uncertainty and inter-instrument bias for the last hour, day, etc. We continuously accumulate the PDFs for each sensor over a variety of time scales and compare it to its nearest neighbors within a neighborhood radius. Any calibration drift in a sensor will be quickly identified as part of the fully automated, real-time workflow, wherein we will automatically be comparing each sensor's PDFs to its neighbor's PDFs, and to the reference instrument's PDFs. As each sensor is in a slightly different local environment, the sensor bias drift for each sensor will be different.

1.3. Characterizing the Temporal and Spatial Scales of Urban Air Pollution

This study focused on the calibration of low cost sensors is part of a larger endeavor with the goal of characterizing the temporal and spatial scales of urban pollution. The temporal and spatial scales of each atmospheric component are intimately connected. The resolution used in atmospheric chemistry modeling tools is often driven by the computational resources available. The spatial resolution of observational networks is often determined by the fiscal resources available. It is worth taking a step back and characterizing what the actual spatial scales are for each chemical component of urban atmospheric chemistry. Based on our street level surveys providing data at resolution higher than one meter, it is clear that the spatial scales are dependent on several factors—the synoptic situation, the distribution of sources, the terrain, etc. In the larger study we characterized the spatial scales of multi-specie urban pollution by using a hierarchy of measurement capabilities that include: (1) A zero emission electric survey vehicle with comprehensive gas, particulate, irradiance, and ionizing radiation sensing. (2) An ensemble of more than one hundred street level sensors making measurements every few seconds of a variety of gases, and of particulates, light levels, temperature, pressure, and humidity. Each sensor is accurately calibrated against a reference standard using machine learning. This paper documents an example of low-cost sensor calibration for airborne particulate observations.

1.4. Societal Relevance

What are the characteristic spatial scales of each chemical species and how does this depend on issues such as the synoptic situation? These are basic questions that are helpful to quantify when considering atmospheric chemistry; when looking forward to the next generation of modeling tools and

observing systems (whether from space or ground-based networks); and when evaluating mitigation strategies, especially with regard to co-benefits for air pollution and greenhouse gas reduction and investigating the evolution of urban air composition in a warming climate. To be able to quantify these spatial and temporal scales we need a comprehensive observing system, so being able to use low cost sensors is of great assistance to achieving this goal.

The Dallas–Fort Worth (DFW) metroplex (where our study was conducted) is the largest inland urban area in the United States and the nation's fourth largest metropolitan area. Nearly a third of Texans, more than seven million inhabitants, live in the DFW area. A population which is growing by a thousand people every day. DFW is an area with an interesting variety of specific pollution sources with unique signatures that can provide a useful testbed for generalizing a measurement strategy for dense urban environments. For more than two decades the DFW area has been in continuous violation of the Clean Air Act. DFW will be one of only ten non-California metropolitan areas still in violation of the Clean Air Act in 2025 unless major changes take place. This has already had a detrimental health impact; e.g., even though the average childhood asthma rate is 7% in Texas, and the national average is 9%, the DFW childhood asthma rate is 20%–25%. Second only to the Northeast, DFW ranks second in the number of annual deaths due to smog. Further, a leading factor in poor learning outcomes in high-schools is absenteeism, a leading cause of absenteeism is asthma, and key trigger for asthma is airborne pollution [21]. Physical exertion in the presence of high pollution levels is more likely to lead to an asthmatic event. The sensors calibrated in this study were provided to high schools and high school coaches so that simple, practical decisions can be made to reduce adverse health outcomes; e.g., given the levels of pollen/pollution today, should physical education/practice be outside or inside?

2. The Datasets Used

All of the measurements were made at our own field calibration station in an ambient environment. The calibration of the low cost AphaSense OPC occurred prior to their deployment across the dense urban environment of DFW. In this study we used machine learning to bring together two distinct types of data. First, we used accurate in-situ observations made by a research grade particulate spectrometer. Second, we used observations from inexpensive optical particle counters. The inexpensive sensors are particularly useful as they can be readily deployed at scale.

2.1. Research Grade Optical Particle Counter

The particulate spectrometer is a laser based Optical Particle Counter (OPC). In this study we used a GRIMM Laser Aerosol Spectrometer and Dust Monitor Model 1.109 (Germany). The sensor has the capability of measuring particulates of diameters between 0.25 and 32 μm distributed within 32 size channels. Such a wide range of diameter space is made possible due to intensity modulation of the laser source. Particulates pumped into the sensor are detected through scattering a laser beam of 655 nm into a light trap. The laser beam is aimed at particulates coming through a sensing chamber at a flow-rate of 1.21 L/min. The device classifies particulates into specific size classes subject to its intensity [22]. The optical arrangement of the sensor is staged such that a curved optical mirror placed at an average scattering angle of 90° collects and redirects the scattered light towards a photo sensor. The wide angle of the optical mirror (120°) is meant to increase the light intensity redirected towards the photo sensor within the Rayleigh scattering domain which decreases the minimum detectable particle size. Furthermore, it compensates for Mie Scattering undulations caused by monochromatic illumination. The sensing period of the GRIMM sensor was set to 6 s, and for each time window provided three standardized mass fractions; namely, based on occupational health (repairable, thoracic, and alveolic) according to EN 481, and PM_1, $PM_{2.5}$, and PM_{10}.

2.2. Low Cost Optical Particle Counters

There are several readily available optical particle counters (OPC) which are useful, but much less accurate compared to research grade sensors. In this study, we focus on using such sensors,

together with machine learning, to get as close as possible to the accuracy of research grade PM sensors. After the application of the machine learning calibration, these lower cost sensors perform admirably. In order for low cost sensors to provide an improved picture of PM levels, a careful calibration is required. The current study used an Alpha Sense OPC-N3 (http://www.alphasense.com/) together with a cheaper environmental sensor (Bosch BME280) as data collectors. The OPC-N3 is compact (75 mm × 60 mm × 65 mm) in size and weighs under 105 g, but uses similar technology to the conventional OPCs where particle size is determined via a calibration based on Mie scattering. Unlike most OPCs the OPC-N3 does not include a pump and a replaceable particle filter in order to pump aerosol samples through a narrow inlet tube; hence, avoiding the need for regular maintenance. A sufficient airflow through the sensor is made possible with a low powered micro fan producing a sample flow rate of 280 mL/min. The OPC-N3 is capable of on-board data logging and measuring particulates with diameters up to 40 µm. This enables the OPC-N3 to measure pollen and other biological particulates. The on-board data is saved within an SD card which can be accessed through micro-USB cable connected to the OPC. Furthermore, the OPC-N3's lower sensing diameter is 0.35 µm, as opposed to its predecessor's (OPC-N2) limit of 0.38 µm. The wider range of sensing is made possible via the OPC switching between high and low gain modes automatically. The OPC-N3 calculates its PM values using the method defined by the European Standard EN 481 [23].

2.3. Caveat: Particulate Refractive Index

The observations made by optical particle counters are sensitive to the refractive index of the particulates and their light absorbing properties. The retrieved size distributions and the mass-concentrations can be biased, depending on the nature of the particulates. The current study did not explore the accuracy implications of this. A future study is underway which includes direct measurements of black carbon that will allow us to begin to explore these aspects. The machine learning paradigm is readily extensible to include these aspects, even though not explicitly addressed in this study. Machine learning is an ideal approach for the calibration of lower cost optical particle counters.

3. Machine Learning

Machine learning has already proved useful in a wide variety of applications in science, business, health care, and engineering. Machine learning allows us to *learn by example*, and to *give our data a voice*. It is particularly useful for those applications for which we do *not* have a complete theory, yet which are of significance. Machine learning is an automated implementation of the scientific method [24], following the same process of generating, testing, and discarding or refining hypotheses. While a scientist or engineer may spend their entire career coming up with and testing a few hundred hypotheses, a machine-learning system can do the same in a fraction of a second. Machine learning provides an objective set of tools for automating discovery. It is therefore not surprising that machine learning is currently revolutionizing many areas of science, technology, business, and medicine [25,26].

Machine learning is now being routinely used to work with large volumes of data in a variety of formats, such as images, videos, sensors, health records, etc. Machine learning can be used in understanding this data and create predictive and classification tools. When machine learning is used for regression, empirical models are built to predict continuous data, facilitating the prediction of future data points, e.g., algorithmic trading and electricity load forecasting. When machine learning is used for classification, empirical models are built to classify the data into different categories, aiding in the more accurate analysis and visualization of the data. Applications of classification include facial recognition, credit scoring, and cancer detection. When machine learning is used for clustering, or unsupervised classification, it aids in finding the natural groupings and patterns in data. Applications of clustering include medical imaging, object recognition, and pattern mining. Object recognition is a process for identifying a specific object in a digital image or video. Object recognition algorithms rely on matching, learning, or pattern recognition algorithms using appearance-based or feature-based techniques.

These technologies are being used for applications such as driver-less cars, automated skin cancer detection, etc.

Machine learning is an automated approach to building empirical models from the data alone. A key advantage of this is that we make no a priori assumptions about the data, its functional form, or its probability distributions. It is an empirical approach. However, it also means that for machine learning to provide the best performance we do need a comprehensive, representative set of examples, that spans as much of the parameter space as possible. This comprehensive set of examples is referred to as the 'training data'.

So, for a successful application of machine learning we have two key ingredients, both of which are essential, a machine learning algorithm, and a comprehensive training data set. Then, once the training has been performed, we should test its efficacy using an independent validation data set to see how well it performs when presented with data that the algorithm has not previously seen; i.e., test its 'generalization'. This can be, for example, a randomly selected subset of the training data that was held back and then utilized for independent validation.

It should be noted, that with a given machine learning algorithm, the performance can go from poor to outstanding with the provision of a progressively more complete training data set. Machine learning really is learning by example, so it is critical to provide as complete a training data set as possible. At times, this can be a labor intensive endeavor.

We have used machine learning in many previous studies [19,21,25–56]. In this study we used machine learning for multivariate non-linear non-parametric regression. Some of the commonly used regression algorithms include neural networks [57–62], support vector machines [63–67], decision trees [68], and ensembles of trees such as random forests [69–71]. Previously we used a similar approach to cross-calibrate satellite instruments [19,25–28]. Recently other studies also used machine learning to calibrate low cost sensors [72,73].

Ensemble Machine Learning

Multiple approaches for non-linear non-parametric machine learning were tried, including neural networks, support vector regression, and ensembles of decision trees. The best performance was found using an ensemble of decision trees with hyper-parameter optimization [68–71]. The specific implementation used was that provided by the Mathworks in the `fitrensemble` function which is part of the Matlab Statistics and Machine Learning Toolbox. Hyperparameter optimization was used so that the optimal choice was made for the following attributes: learning method (bagging or boosting), maximum number of learning cycles, learning rate, minimum leaf size, maximum number of splits, and the number of variables to sample.

There were 72 inputs to our multivariate non-linear non-parametric machine learning regression; these included the particle counts for each of the 24 size bins measured by the OPC-N3; the OPC-N3 estimates of PM_1, $PM_{2.5}$, and PM_{10}; a suite of OPC performance variables, including the reject ratio; and particularly importantly, the ambient atmospheric pressure, temperature, and humidity. The OPC-N3 sensor includes two photo diodes that record voltages which are eventually translated into particle count data. However, particles which are not entirely in the OPC-N3 laser beam, or are passing down the edge, are rejected and this is recorded in the "reject ratio" parameter. This leads to better sizing of particles, and hence plays an important role within the machine learning calibration.

Each of the six outputs we wished to estimate had its own empirical model. The performances of each of these six models in their independent validations are shown in Figures 1 and 2. The outputs we estimated were the six variables measured by the reference instrument, the research grade optical particle counts, namely, of PM_1, $PM_{2.5}$, and PM_{10}; and the standardized occupational health repairable, thoracic, and alveolic mass fractions. The alveolic fraction is the mass fraction of inhaled particles penetrating to the alveolar region (maximum deposition of particles with a size $\approx$2 μm). The Thoracic fraction is the mass fraction of inhaled particles penetrating beyond the larynx (<10 μm). The respirable fraction is the mass fraction of inhaled particles penetrating to the unciliated airways (<4 μm).

The inhalable fraction is the mass fraction of total airborne particles which is inhaled through the nose and mouth (<20 μm). For each of these six parameters we created an empirical multivariate non-linear non-parametric machine learning regression model with hyper-parameter optimization.

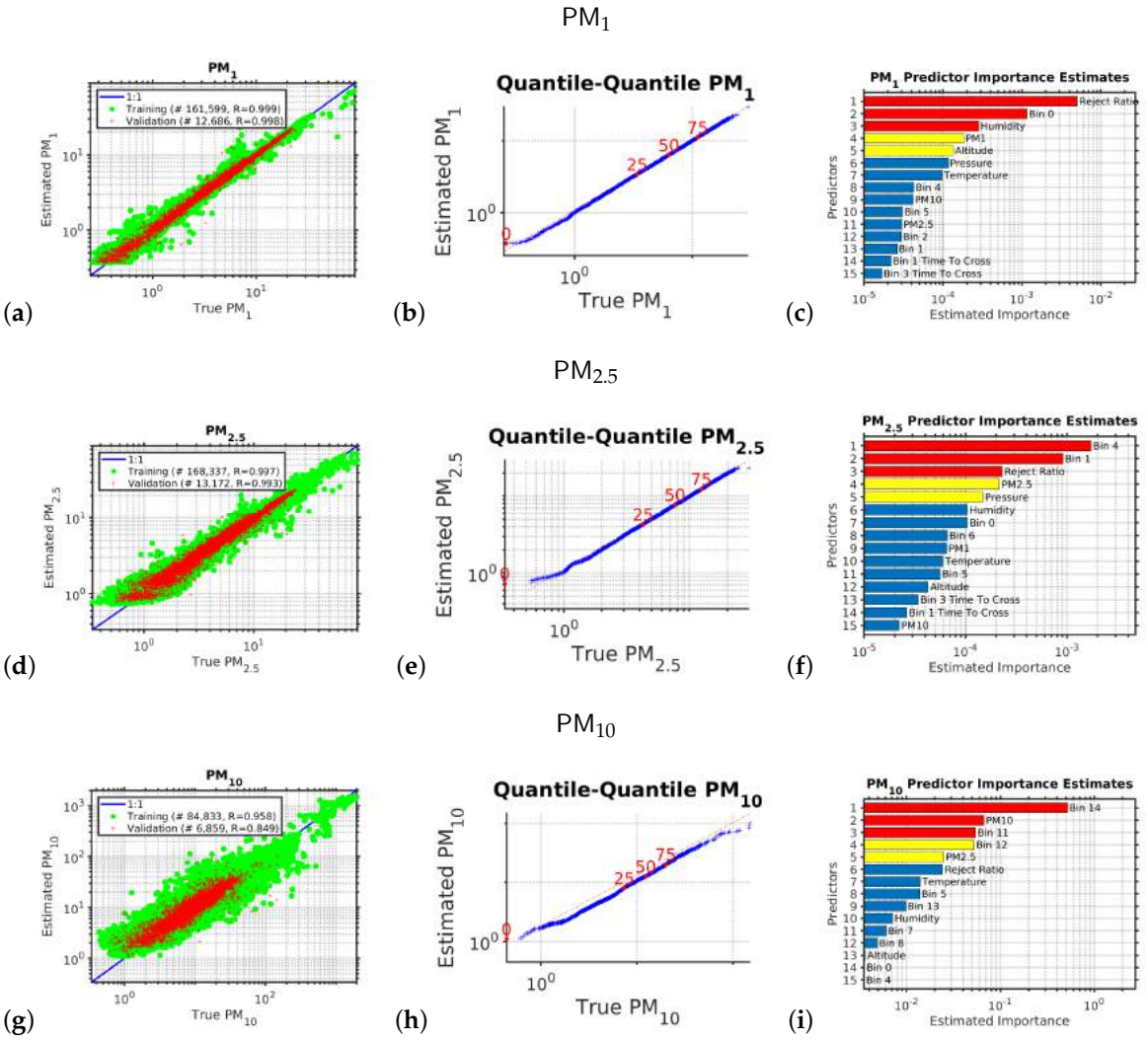


Figure 1. This figure shows the results of the multivariate non-linear non-parametric machine learning regression for particulate matter PM_1 (panels (**a**)–(**c**)), $PM_{2.5}$ (panels (**d**)–(**f**)), and PM_{10} (panels (**g**)–(**i**)). The left hand column of plots shows the log–log axis scatter diagrams with the x-axis showing the PM abundance from the expensive reference instrument and the y-axis showing the PM abundance provided by calibrating the low-cost instrument using machine learning. The green circles are the training data; the red pluses are the independent validation dataset. The blue line shows the ideal response. The middle column of plots shows the quantile–quantile plots for the machine learning validation data, with the x-axis showing the percentiles from the probability distribution function of the PM abundance from the expensive reference instrument and the y-axis showing the percentiles from the probability distribution function of the estimated PM abundance provided by calibrating the low-cost instrument using machine learning. The dotted red line shows the ideal response. The right hand column of plots shows the relative importance of the input variables for calibrating the low-cost optical particle counters using machine learning.

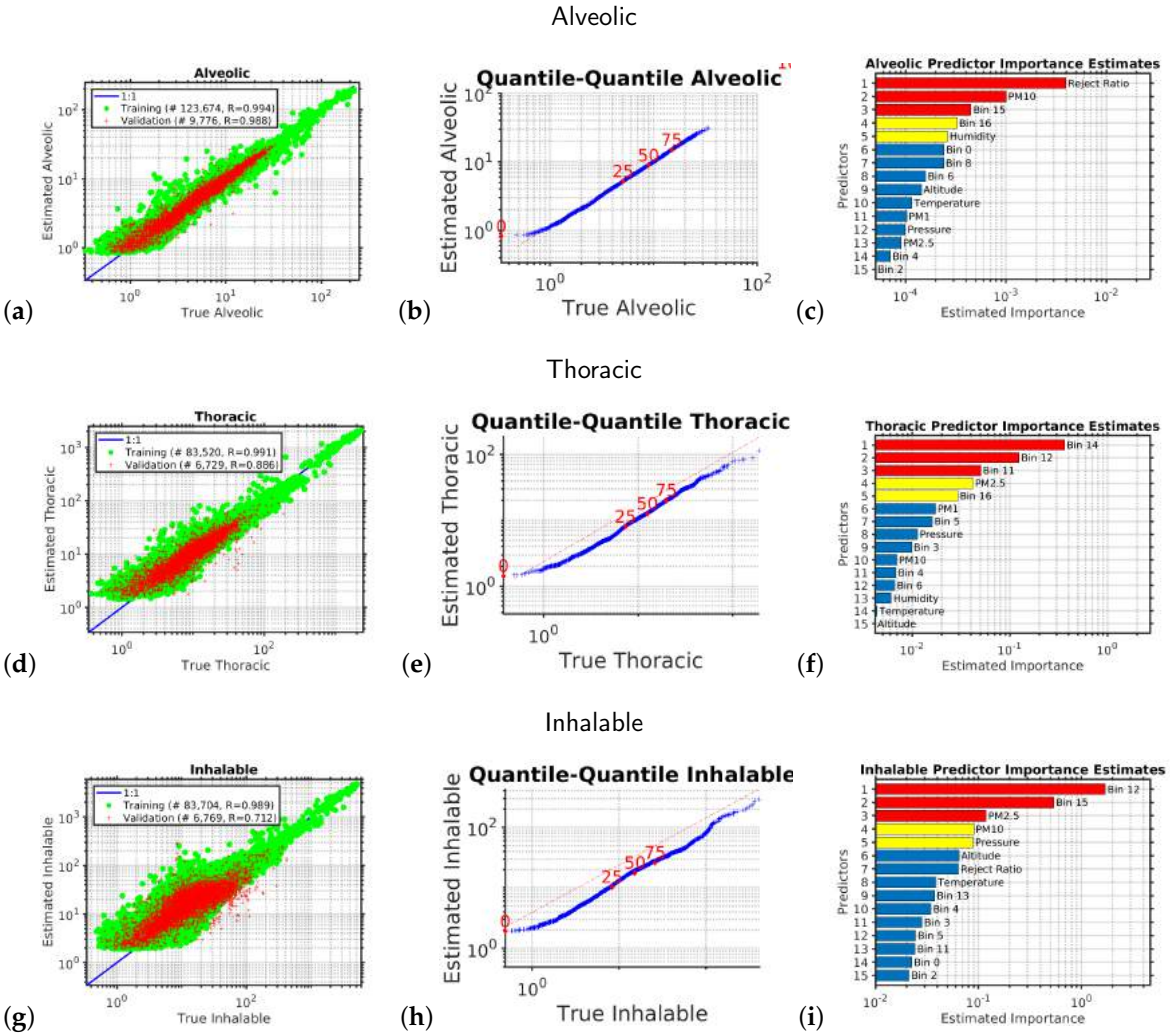


Figure 2. This figure shows the results of the multivariate non-linear non-parametric machine learning regression for the alveolic (panels (**a**)–(**c**)), thoracic (panels (**d**)–(**f**)), and inhalable size fractions (panels (**g**–**i**)). The left hand column of plots shows the log–log axis scatter diagrams with the x-axis showing the PM abundance from the expensive reference instrument and the y-axis showing the PM abundance provided by calibrating the low-cost instrument using machine learning. The green circles are the training data; the red pluses are the independent validation dataset. The blue line shows the ideal response. The middle column of plots shows the quantile–quantile plots for the machine learning validation data, with the x-axis showing the percentiles from the probability distribution function of the PM abundance from the expensive reference instrument and the y-axis showing the percentiles from the probability distribution function of the estimated PM abundance provided by calibrating the low-cost instrument using machine learning. The dotted red line shows the ideal response. The right hand column of plots shows the relative importance of the input variables for calibrating the low-cost optical particle counters using machine learning.

4. Results

Calibrating the Low Cost Optical Particle Counters Using Machine Learning

Figure 1 shows the the results of the multivariate non-linear non-parametric machine learning regression for PM_1 (panels a to c), $PM_{2.5}$ (panels d to f), and PM_{10} (panels g to i). The left hand column of plots shows the log–log axis scatter diagrams with the *x*-axis showing the PM abundance from the

expensive reference instrument and the y-axis showing the PM abundance provided by calibrating the low-cost instrument using machine learning.

For the left hand column of plots in Figure 1 (the scatter diagrams), for a perfect calibration, the scatter plot would be a straight line with a slope of one and a y-axis intercept of zero; the blue line shows the ideal response. We can see that multivariate non-linear non-parametric machine learning regression that we used in this study employing an ensemble of decision trees with hyper-parameter optimization performed very well (panels a, d, and g). In each scatter diagram the green circles are the data used to train the ensemble of decision trees; the red pluses are the independent validation data used to test the generalization of the machine learning model.

We can see that the performance is best for the smaller particles that stay lofted in the air for a long period and do not rapidly sediment, so when comparing the scatter diagram correlation coefficients, r, for the independent validation test data (red-points) we see that $r_{\mathrm{PM}_1} > r_{\mathrm{PM}_{2.5}} > r_{\mathrm{PM}_{10}}$.

For the middle column of plots in Figure 1 (the quantile–quantile plots), we are comparing the *shape* of the probability distribution (PDF) of all the PM abundance data collected by the expensive reference instrument to that of the the PM abundance provided by calibrating the low-cost instrument using machine learning. A $\log_{10}$ scale is used with a tick mark every decade. The dotted red line in each quantile–quantile plot shows the ideal response. The red numbers indicate the percentiles (0, 25, 50, 75, 100). If the quantile–quantile plot is a straight line, that means both PDFs have *exactly* the same shape, as we are plotting the percentiles of one PDF against the percentiles of the other PDF. Usually, we would like to see a straight line at least between the 25th and 75th percentiles; in this case, we have a straight line over the entire PDF, which demonstrates that the machine learning calibration performed well.

The right hand column of plots shows the relative importance of the input variables for calibrating the low-cost optical particle counters using machine learning. The relative importance metric is a measure of the error that results if that input variable is omitted. In the right hand column of bar plots we have sorted the importance metrics into descending order, so the variable represented by the uppermost bar in each each case was the most important variable for performing the calibration; the second bar was the second most important; etc. We note that along with the number of particles counted in each size bin, it is important to measure the temperature, pressure, and humidity to be able to accurately calibrate the low cost OPC against the reference instrument. The data also suggests that the parameter "reject ratio" carries a greater deal of importance with respect to the calibration. OPC-N3 comprises two photo diodes which records voltages which are eventually translated into particle count data. However, particles which are not entirely in the beam or are passing down the edge are rejected and that is reflected on the parameter "reject ratio". This leads to better sizing of particles, and hence plays a vital role within the ML calibration.

Another division of occupational health based size-selective sampling is defined by assessing the subset of particles that can reach a selective region of the respiratory system. On this basis three main fractions were defined: inhalable, thoracic, and respirable [74–76]. Studies have shown that exposure of excess particulate matter has alarming negative health effects [77]. The smallest sizes of particulate matter are capable of penetrating through to the lungs or even to one's blood stream.

Figure 2 is similar to Figure 1 and shows the results of the multivariate non-linear non-parametric machine learning regression for the alveolic, thoracic, and inhalable size fractions. As would be expected, we see that the performance is best for the smaller particles that stay lofted in the air for a long period and do not rapidly sediment, so when comparing the scatter diagram correlation coefficients, r, for the independent validation test data (red-points) we see that $r_{Alveolic} > r_{Thoracic} > r_{Inhalable}$.

5. Operational Use of the Calibration and Periodic Validation Updates

The calibration just described occurred pre-deployment of the sensors into the dense urban environment. Once these initial field calibration measurements were made over a period of several months, in the manner described above, the multi-variate non-linear non-parametric empirical machine

learning model was applied in real time to the live stream of observations coming from each of our air quality sensors deployed across the dense urban environment of the Dallas Fort Worth metroplex. These corrected measurements were then made publicly available as open data and depicted on a live map and dashboard.

Building in continual calibration to a network of sensors will enable long-term, consistent, and reliable data. While much effort has been recently placed on the connectivity of large disbursed IoT networks, little to no effort has been spent on the automated calibration, bias-detection, and uncertainty estimation necessary to make sure the information collected is sound. This is one of our primary goals. This is based on extensive previous work funded by NASA for satellite validation.

After deployment, a zero emission electric car carrying our reference was used, to routinely drive past all the deployed sensors to provide ongoing routine calibration and validation. An electric vehicle does not contribute any ambient emissions, and so, is an ideal mobile platform for our reference instruments.

For optimal performance, the implementation combines edge and cloud computing. Each sensor node takes a measurement at least every 10 s. The observations are continually time-stamped at the nodes and streamed to our cloud server, the central server aggregating all the data from the nodes, and managing them. To prevent data loss, the sensor nodes store any values that have not been transmitted to the cloud server for reasons, including communication interruptions, in a persistent buffer. The local buffer is emptied to the cloud server at the next available opportunity.

Data from all sensors are archived and serve as an open dataset that can be publicly accessed. The observed probability distribution functions (PDFs) from each sensor are automatically compared in real time to the PDFs from the neighboring sensors within a neighborhood radius. These neighborhood sensors include measurements from the electric car/mobile validation sensors. This comparison was used to estimate the size resolved measurement uncertainty and size resolved inter-instrument bias for the last hour, day, week, month, and year. We continuously accumulated the PDF for each sensor over a variety of time scales (h, day, week, month, and year) and compare it to its nearest neighbors within a neighborhood radius.

Any calibration drift in a sensor will be quickly identified as part of a fully automated real-time workflow, where we will automatically be comparing each sensor's PDFs to its neighbor's PDFs, and to the reference instruments' PDFs. As each sensor is in a slightly different local environment, the sensor bias drift for each sensor will be different. We have previously shown that machine learning can be used to effectively correct these inter-sensor biases [19]. As a result, the overall distributed sensing system will not just be better characterized in terms of its uncertainty and bias, but provide improved measurement stability over time.

6. Conclusions

We have shown that machine learning can be used to effectively calibrate lower cost optical particle counters. For this calibration it is critical that measurements of the atmospheric pressure, humidity, and temperature are included. Once the machine learning calibration was applied to the low cost sensors, independent validation using scatter diagrams and quantile–quantile plots showed that, not only was the calibration effective, but the shape of the resulting probability distribution of observations was very well preserved.

These low cost sensors are being deployed at scale across the dense urban environment of the Dallas Fort Worth metroplex for characterizing both the temporal and spatial scales of urban air pollution and for providing high schools and high school coaches a tool to assist in making better decisions to reduce adverse health outcomes; e.g., given the levels of pollen/pollution today should physical education/practice be outside or inside?

Author Contributions: conceptualization, D.J.L.; sensor construction, L.O.H.W., D.R.K. and A.R.A.; sensor calibration and validation, L.O.H.W.; methodology, L.O.H.W. and D.J.L.; software, L.O.H.W. and D.J.L.; formal analysis, L.O.H.W. and D.J.L.; investigation, L.O.H.W. and D.J.L.; resources, D.J.L.; data curation, L.O.H.W.; writing—original draft preparation, L.O.H.W. and D.J.L.; writing—review and editing, L.O.H.W., S.T., and D.J.L.; visualization, L.O.H.W. and D.J.L.; supervision, D.J.L.; project administration, D.J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by USAMRMC, award number W81XWH-18-1-0400. The National Science Foundation CNS Division of Computer and Network Systems, grant 1541227. Earth Day Texas. Downwinders at Risk. The City of Plano, TX.

Acknowledgments: We warmly thank the anonymous reviewers for their suggestions to improve this paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

OPC Optical Particle Counter
PDF Probability Distribution Function
PM Particulate Matter

References

1. Boucher, O. *Atmospheric Aerosols: Properties and Climate Impacts*; Springer: Haarlem, The Netherlands, 2015.
2. Charlson, R.J.; Schwartz, S.; Hales, J.; Cess, R.D.; Coakley, J.J.; Hansen, J.; Hofmann, D. Climate forcing by anthropogenic aerosols. *Science* **1992**, *255*, 423–430. [CrossRef]
3. Ramanathan, V.; Crutzen, P.; Kiehl, J.; Rosenfeld, D. Aerosols, climate, and the hydrological cycle. *Science* **2001**, *294*, 2119–2124. [CrossRef] [PubMed]
4. Dubovik, O.; Holben, B.; Eck, T.F.; Smirnov, A.; Kaufman, Y.J.; King, M.D.; Tanré, D.; Slutsker, I. Variability of absorption and optical properties of key aerosol types observed in worldwide locations. *J. Atmos. Sci.* **2002**, *59*, 590–608. [CrossRef]
5. Guenther, A.; Karl, T.; Harley, P.; Wiedinmyer, C.; Palmer, P.; Geron, C. Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature). *Atmos. Chem. Phys.* **2006**, *6*, 3181–3210. [CrossRef]
6. Hallquist, M.; Wenger, J.C.; Baltensperger, U.; Rudich, Y.; Simpson, D.; Claeys, M.; Dommen, J.; Donahue, N.; George, C.; Goldstein, A.; et al. The formation, properties and impact of secondary organic aerosol: Current and emerging issues. *Atmos. Chem. Phys.* **2009**, *9*, 5155–5236. [CrossRef]
7. Kanakidou, M.; Seinfeld, J.; Pandis, S.; Barnes, I.; Dentener, F.; Facchini, M.; Dingenen, R.V.; Ervens, B.; Nenes, A.; Nielsen, C.; et al. Organic aerosol and global climate modelling: A review. *Atmos. Chem. Phys.* **2005**, *5*, 1053–1123. [CrossRef]
8. Allen, M.R.; Barros, V.R.; Broome, J.; Cramer, W.; Christ, R.; Church, J.A.; Clarke, L.; Dahe, Q.; Dasgupta, P.; Dubash, N.K.; et al. *IPCC Fifth Assessment Synthesis Report-Climate Change 2014 Synthesis Report*; IPCC: Geneva, Switzerland, 2014.
9. Dockery, D.W.; Pope, C.A.; Xu, X.; Spengler, J.D.; Ware, J.H.; Fay, M.E.; Ferris, B.G., Jr.; Speizer, F.E. An association between air-pollution and mortality in 6 United-States cities. *N. Engl. J. Med.* **1993**, *329*, 1753–1759. [CrossRef]
10. Oberdörster, G.; Oberdörster, E.; Oberdörster, J. Nanotoxicology: An emerging discipline evolving from studies of ultrafine particles. *Environ. Health Perspect.* **2005**, *113*, 823–839. [CrossRef]
11. Pope, C.A., III; Burnett, R.T.; Thun, M.J.; Calle, E.E.; Krewski, D.; Ito, K.; Thurston, G.D. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA* **2002**, *287*, 1132–1141. [CrossRef]
12. Pope, C.A., III; Dockery, D.W. Health effects of fine particulate air pollution: Lines that connect. *J. Air Waste Manag. Assoc.* **2006**, *56*, 709–742. [CrossRef]
13. Cheng, M.; Liu, W. *Airborne Particulates*; Nova Science Publishers: Hauppauge, NY, USA, 2009.
14. Chin, M. *Atmospheric Aerosol Properties and Climate Impacts*; DIANE Publishing Company: Collingdale, PA, USA, 2009.

15. Lim, S.S.; Vos, T.; Flaxman, A.D.; Danaei, G.; Shibuya, K.; Adair-Rohani, H.; AlMazroa, M.A.; Amann, M.; Anderson, H.R.; Andrews, K.G.; et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **2012**, *380*, 2224–2260. [CrossRef]
16. Stocker, T. *Climate Change 2013: The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*; Cambridge University Press: Cambridge, UK, 2014.
17. Seinfeld, J. *Atmospheric Chemistry and Physics of Air Pollution*; Wiley: Hoboken, NJ, USA, 1986.
18. Pöschl, U. Atmospheric aerosols: Composition, transformation, climate and health effects. *Angew. Chem. Int. Ed.* **2005**, *44*, 7520–7540. [CrossRef] [PubMed]
19. Lary, D.J.; Remer, L.A.; MacNeill, D.; Roscoe, B.; Paradise, S. Machine Learning and Bias Correction of MODIS Aerosol Optical Depth. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 694–698. [CrossRef]
20. Lary, D.J. Representativeness uncertainty in chemical data assimilation highlight mixing barriers. *Atmos. Sci. Lett.* **2004**, *5*, 35–41. [CrossRef]
21. Lary, M.A.; Allsop, L.; Lary, D.J. Using Machine Learning to Examine the Relationship Between Asthma and Absenteeism. *Environ. Model. Assess.* **2019**, *191*, 332. [CrossRef] [PubMed]
22. Broich, A.V.; Gerharz, L.E.; Klemm, O. Personal monitoring of exposure to particulate matter with a high temporal resolution. *Environ. Sci. Pollut. Res.* **2012**, *19*, 2959–2972. [CrossRef]
23. Alphasense. *Alphasense User Manual OPC-N3 Optical Particle Counter*; Alphasense Ltd.: Great Notley, UK, 2018.
24. Domingos, P. *The Master Algorithm: How the Quest for the Ultimate Learning Machine will Remake Our World*; Basic Books: New York, NY, USA, 2015.
25. Lary, D.J.; Alavi, A.H.; Gandomi, A.H.; Walker, A.L. Machine learning in geosciences and remote sensing. *Geosci. Front.* **2016**, *7*, 3–10. [CrossRef]
26. Lary, D.J.; Zewdie, G.K.; Liu, X.; Wu, D.; Levetin, E.; Allee, R.J.; Malakar, N.; Walker, A.; Mussa, H.; Mannino, A.; et al. Machine learning applications for earth observation. In *Earth Observation Open Science and Innovation*; Springer: Berlin, Germany, 2018; Volume 165.
27. Brown, M.E.; Lary, D.J.; Vrieling, A.; Stathakis, D.; Mussa, H. Neural networks as a tool for constructing continuous NDVI time series from AVHRR and MODIS. *Int. J. Remote Sens.* **2008**, *29*, 7141–7158. [CrossRef]
28. Lary, D.; Aulov, O. Space-based measurements of HCl: Intercomparison and historical context. *J. Geophys. Res. Atmos.* **2008**, *113*. [CrossRef]
29. Lary, D.; Müller, M.; Mussa, H. Using neural networks to describe tracer correlations. *Atmos. Chem. Phys.* **2004**, *4*, 143–146. [CrossRef]
30. Malakar, N.; Lary, D.; Gencaga, D.; Albayrak, A.; Wei, J. Towards identification of relevant variables in the observed aerosol optical depth bias between MODIS and AERONET observations. In *AIP Conference Proceedings*; AIP: Clermont-Ferrand, France, 2013; Volume 1553, pp. 69–76.
31. Lary, D.J. *Artificial Intelligence in Geoscience and Remote Sensing*; INTECH Open Access Publisher: London, UK, 2010.
32. Malakar, N.K.; Lary, D.J.; Moore, A.; Gencaga, D.; Roscoe, B.; Albayrak, A.; Wei, J. Estimation and bias correction of aerosol abundance using data-driven machine learning and remote sensing. In Proceedings of the 2012 Conference on Intelligent Data Understanding, Boulder, CO, USA, 24–26 October 2012; pp. 24–30.
33. Lary, D.J. Using Multiple Big Datasets and Machine Learning to Produce a New Global Particulate Dataset: A Technology Challenge Case Study. In *AGU Fall Meeting Abstracts*; American Geophysical Union: Washington, DC, USA, 2013.
34. Lary, D. Using Neural Networks for Instrument Cross-Calibration. In *AGU Fall Meeting Abstracts*; American Geophysical Union: Washington, DC, USA, 2007.
35. Albayrak, A.; Wei, J.; Petrenko, M.; Lary, D.; Leptoukh, G. MODIS Aerosol Optical Depth Bias Adjustment Using Machine Learning Algorithms. In *AGU Fall Meeting Abstracts*; American Geophysical Union: Washington, DC, USA, 2011.
36. Brown, M.; Lary, D.; Mussa, H. Using Neural Nets to Derive Sensor-Independent Climate Quality Vegetation Data based on AVHRR, SPOT-Vegetation, SeaWiFS and MODIS. In *AGU Spring Meeting Abstracts*; American Geophysical Union: Washington, DC, USA, 2006.
37. Lary, D.; Müller, M.; Mussa, H. Using neural networks to describe tracer correlations. *Atmos. Chem. Phys. Discuss.* **2003**, *3*, 5711–5724. [CrossRef]

38. Malakar, N.; Lary, D.; Allee, R.; Gould, R.; Ko, D. Towards Automated Ecosystem-based Management: A case study of Northern Gulf of Mexico Water. In *AGU Fall Meeting Abstracts*; American Geophysical Union: Washington, DC, USA, 2012.
39. Lary, D.J. BigData and Machine Learning for Public Health. In Proceedings of the 142nd APHA Annual Meeting and Exposition 2014, New Orleans, LA, USA, 15–19 November 2014.
40. Lary, D.; Lary, T.; Sattler, B. Using Machine Learning to Estimate Global PM2.5 for Environmental Health Studies. *Environ. Health Insights* **2015**, *9*, 41. [CrossRef] [PubMed]
41. Kneen, M.A.; Lary, D.J.; Harrison, W.A.; Annegarn, H.J.; Brikowski, T.H. Interpretation of satellite retrievals of PM2.5 over the Southern African Interior. *Atmos. Environ.* **2016**, *128*, 53–64. [CrossRef]
42. Lary, D.; Nikitkov, A.; Stone, D.; Nikitkov, A. Which Machine-Learning Models Best Predict Online Auction Seller Deception Risk? Available online: https://davidlary.info/wp-content/uploads/2012/08/2010-AAA-Strategic-and-Emerging-Technologies.pdf (accessed on 22 December 2019).
43. Medvedev, I.R.; Schueler, R.; Thomas, J.; Kenneth, O.; Nam, H.J.; Sharma, N.; Zhong, Q.; Lary, D.J.; Raskin, P. Analysis of exhaled human breath via terahertz molecular spectroscopy. In Proceedings of the 2016 41st International Conference on Infrared, Millimeter, and Terahertz waves (IRMMW-THz), Copenhagen, Denmark, 25–30 September 2016; pp. 1–2.
44. Lary, D.J.; Lary, T.; Sattler, B. Using Machine Learning to Estimate Global Particulate Matter for Environmental Health Studies. *Geoinform. Geostat. Overv.* **2016**, *4*. [CrossRef]
45. Zhong, Q.; Sharma, N.; Choi, W.; Schueler, R.; Medvedev, I.R.; Nam, H.J.; Raskin, P.; De Lucia, F.C.; McMillan, J.P. *Demonstration of Breath Analyses Using CMOS Integrated Circuits for Rotational Spectroscopy*; International Workshop on Nanodevice Technologies: Hiroshima, Japan, 2017.
46. Wu, D.; Zewdie, G.K.; Liu, X.; Kneed, M.; Lary, D.J. Insights Into the Morphology of the East Asia PM2.5 Annual Cycle Provided by Machine Learning. *Environ. Health Insights* **2017**, *11*, 1–7. [CrossRef]
47. Nathan, B.J.; Lary, D.J. Combining Domain Filling with a Self-Organizing Map to Analyze Multi-Species Hydrocarbon Signatures on a Regional Scale. *Environ. Model. Assess.* **2019**, *191*, 337. [CrossRef]
48. Wu, D.; Lary, D.J.; Zewdie, G.K.; Liu, X. Using Machine Learning to Understand the Temporal Morphology of the PM2.5 annual cycle in East Asia. *Environ. Monit. Assess.* **2019**, *191*, 272. [CrossRef]
49. Alavi, A.H.; Gandomi, A.H.; Lary, D.J. Progress of Machine Learning in Geosciences. *Geosci. Front.* **2016**, *7*, 1–2. [CrossRef]
50. Ahmad, Z.; Choi, W.; Sharma, N.; Zhang, J.; Zhong, Q.; Kim, D.Y.; Chen, Z.; Zhang, Y.; Han, R.; Shim, D.; et al. Devices and circuits in CMOS for THz applications. In Proceedings of the 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 3–7 December 2016.
51. Zewdie, G.; Lary, D.J. Applying Machine Learning to Estimate Allergic Pollen Using Environmental, Land Surface and NEXRAD radar Parameters. In *AGU Fall Meeting Abstracts*; American Geophysical Union: Washington, DC, USA, 2018.
52. Malakar, N.K.; Lary, D.; Gross, B. Case Studies of Applying Machine Learning to Physical Observation. In *AGU Fall Meeting Abstracts*; American Geophysical Union: Washington, DC, USA, 2018.
53. Zewdie, G.K.; Lary, D.J.; Levetin, E.; Garuma, G.F. Applying Deep Neural Networks and Ensemble Machine Learning Methods to Forecast Airborne Ambrosia Pollen. *Int. J. Environ. Res. Public Health* **2019**, *16*, 1992. [CrossRef]
54. Zewdie, G.K.; Lary, D.J.; Liu, X.; Wu, D.; Levetin, E. Estimating the daily pollen concentration in the atmosphere using machine learning and NEXRAD weather radar data. *Environ. Monit. Assess.* **2019**, *191*, 418. [CrossRef]
55. Chang, H.H.; Pan, A.; Lary, D.J.; Waller, L.A.; Zhang, L.; Brackin, B.T.; Finley, R.W.; Faruque, F.S. Time-series analysis of satellite-derived fine particulate matter pollution and asthma morbidity in Jackson, MS. *Environ. Monit. Assess.* **2019**, *191*, 280. [CrossRef] [PubMed]
56. Choi, W.; Zhong, Q.; Sharma, N.; Zhang, Y.; Han, R.; Ahmad, Z.; Kim, D.Y.; Kshattry, S.; Medvedev, I.R.; Lary, D.J.; et al. Opening Terahertz for Everyday Applications. *IEEE Commun. Mag.* **2019**, *57*, 70–76.
57. McCulloch, W.; Pitts, W. A Logical calculus of the Ideas Immanent in Nervous Activity. *Bull. Math. Biophys.* **1943**, *5*, 115. [CrossRef]
58. Haykin, S.S. *Kalman Filtering and Neural Networks*; Adaptive and Learning Systems for Signal Processing, Communications, and Control; Wiley: New York, NY, USA, 2001.

59. Haykin, S.S. *New Directions in Statistical Signal Processing: From Systems to Brain*; Neural Information Processing Series; MIT Press: Cambridge, MA, USA, 2007.
60. Haykin, S.S. *Neural Networks: A Comprehensive Foundation*; Macmillan: New York, NY, USA, 1994.
61. Demuth, H.B.; Beale, M.H.; De Jess, O.; Hagan, M.T. *Neural Network Design*, 2nd ed.; Martin Hagan: Notre Dame, IN, USA, 2014.
62. Bishop, C.M. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, UK, 1995.
63. Vapnik, V.N. *Estimation of Dependences Based on Empirical Data*; Springer Series in Statistics; Springer: New York, NY, USA, 1982.
64. Vapnik, V.N. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1995.
65. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
66. Vapnik, V.N. *The Nature of Statistical Learning Theory*, 2nd ed.; Statistics for Engineering and Information Science; Springer: New York, NY, USA, 2000.
67. Vapnik, V.N. *Estimation of Dependences Based on Empirical Data*; Springer: New York, NY, USA, 2006.
68. Safavian, S.R.; Landgrebe, D. A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.* **1991**, *21*, 660–674. [CrossRef]
69. Ho, T.K. The Random Subspace Method for Constructing Decision Forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 832–844.
70. Breiman, L. *Classification and Regression Trees*; The Wadsworth Statistics/Probability Series; Wadsworth International Group: Belmont, CA, USA, 1984.
71. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
72. Li, L.; Zheng, Y.; Zhang, L. Demonstration abstract: PiMi air box—A cost-effective sensor for participatory indoor quality monitoring. In Proceedings of the 13th International Symposium on Information Processing in Sensor Networks, Berlin, Germany, 15–17 April 2014; pp. 327–328.
73. Dong, W.; Guan, G.; Chen, Y.; Guo, K.; Gao, Y. Mosaic: Towards City Scale Sensing with Mobile Sensor Networks. In Proceedings of the 2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS), Melbourne, VIC, Australia, 14–17 December 2015; pp. 29–36. [CrossRef]
74. Bickis, U. Hazard prevention and control in the work environment: Airborne dust. *World Health* **1998**, *13*, 16.
75. Hinds, W.C. *Aerosol Technology: Properties, Behavior, and Measurement of Airborne Particles*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
76. Brown, J.S.; Gordon, T.; Price, O.; Asgharian, B. Thoracic and respirable particle definitions for human health risk assessment. *Part. Fibre Toxicol.* **2013**, *10*, 12. [CrossRef]
77. Mannucci, P.M. Air pollution levels and cardiovascular health: Low is not enough. *Eur. J. Prev. Cardiol.* **2017**, 1851–1853. [CrossRef]

sensors

MDPI

Article

Machine Learning for Light Sensor Calibration

Yichao Zhang *, Lakitha O. H. Wijeratne, Shawhin Talebi and David J. Lary *

Hanson Center for Space Sciences, University of Texas at Dallas, Richardson, TX 75080, USA; lhw150030@utdallas.edu (L.O.H.W.); Shawhin.Talebi@utdallas.edu (S.T.)

* Correspondence: yichao.zhang1@utdallas.edu or hi.yichao.zhang@gmail.com (Y.Z.); David.Lary@utdallas.edu (D.J.L)

Abstract: Sunlight incident on the Earth's atmosphere is essential for life, and it is the driving force of a host of photo-chemical and environmental processes, such as the radiative heating of the atmosphere. We report the description and application of a physical methodology relative to how an ensemble of very low-cost sensors (with a total cost of <$20, less than 0.5% of the cost of the reference sensor) can be used to provide wavelength resolved irradiance spectra with a resolution of 1 nm between 360–780 nm by calibrating against a reference sensor using machine learning. These low-cost sensor ensembles are calibrated using machine learning and can effectively reproduce the observations made by an NIST calibrated reference instrument (Konica Minolta CL-500A with a cost of around USD 6000). The correlation coefficient between the reference sensor and the calibrated low-cost sensor ensemble has been optimized to have $R^2 > 0.99$. Both the circuits used and the code have been made publicly available. By accurately calibrating the low-cost sensors, we are able to distribute a large number of low-cost sensors in a neighborhood scale area. It provides unprecedented spatial and temporal insights into the micro-scale variability of the wavelength resolved irradiance, which is relevant for air quality, environmental and agronomy applications.

Keywords: spectrophotometer; light sensor; machine learning; neural networks

Citation: Zhang, Y.; Wijeratne, L.O.H.; Talebi, S.; Lary, D.J. Machine Learning for Light Sensor Calibration. *Sensors* **2021**, *21*, 6259. https://doi.org/10.3390/s21186259

Academic Editor: Stephen Holler

Received: 22 July 2021
Accepted: 9 September 2021
Published: 18 September 2021

1. Introduction

Sunlight incident on the Earth's atmosphere is essential for life, and it is the driving force of a host of photo-chemical and environmental processes (e.g., photosynthesis, photolysis and atmospheric radiative heating). Consequently, models of atmospheric radiative transfer play a key role in modeling atmospheric chemistry and the weather/climate system (e.g., [1–8]). In order to accurately model the surface irradiance, a complete description of both light absorption, light multiple scattering and surface reflection is required in an atmospheric radiative transfer model. For solar zenith angles $> 75°$, this would also need to account for the spherical geometry of the atmosphere [3,4]. The intensity of atmospheric electromagnetic radiation which reaches the Earth's surface is a strong function of wavelength and the vertical profiles of atmospheric composition and temperature. The vertical profiles of temperature, light scatterers and light absorbers determine the extinction due to absorption and scattering as well as thermal emission.

Atmospheric absorption and multiple-scattering of light both have a significant impact on the surface irradiance (Figure 1). Atmospheric radiative transfer considers the energy transfer of electromagnetic radiation through the atmosphere. The intensity of sunlight, as a function of wavelength, is affected by both the gaseous absorption in the UV and visible portion of the spectrum (including O_3, NO_2, NO_3, HONO and HNO_3 [9,10]), by light scattering from air molecules (Rayleigh scattering), from airborne aerosols (Mie scattering) and by thermal emission in the infrared [1,2,11] (Figure 1).

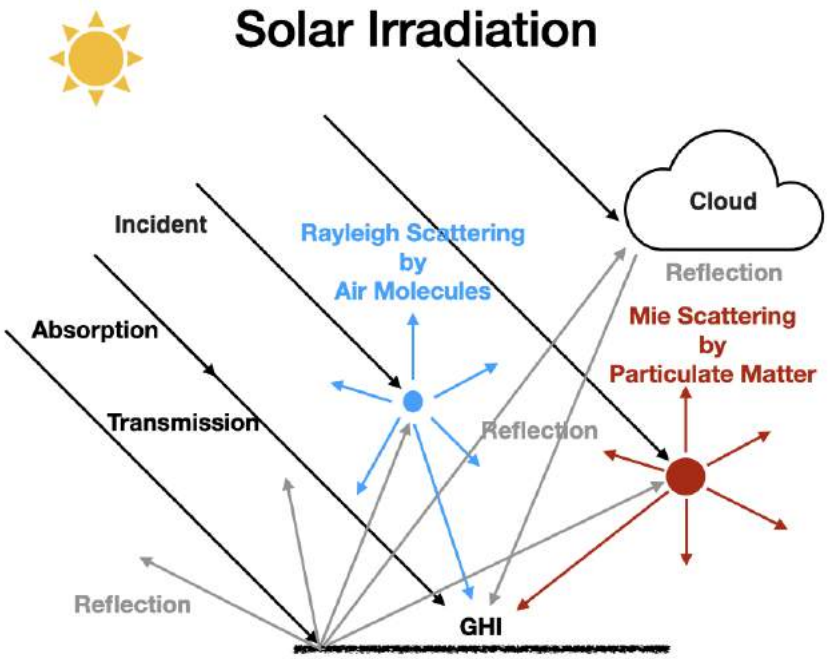


Figure 1. Solar irradiance on the Earth's Surface.

Rayleigh Scattering occurs from gas molecules as the sizes of the molecules are much smaller than the wavelength [12]. The strength of the Rayleigh scattering is proportional to λ^{-4}, where λ is the wavelength of the radiation. Shorter wavelengths scatter more strongly than longer wavelengths, and this is the reason that the sky is blue. Mie scattering occurs when the size of the scatterers is similar to or greater than the wavelength of the light [13]. In the UV and visible portion of the electromagnetic spectrum that we observed in this study (360–780 nm) the main gaseous absorption of light is due to the ozone in a set of different absorption bands (Hartley 200–300 nm, Huggins 310–360 nm, Chappuis 400–650 nm and Wulf in the near infrared).

1.1. Motivation

The goal of this study is to provide an accurately calibrated low-cost wavelength resolved irradiance sensor, which is helpful for biometric pupillometry [14] applications, and is suitable to address the current lack of neighborhood scale real-time solar irradiance data by the provision of very low-cost calibrated measurements. These sensors can be readily deployed at a scale across dense urban environments in order to measure the wavelength resolved irradiance. Sunlight incident on the Earth's atmosphere is essential for life, and it drives atmospheric photo-chemistry, which is central to understanding urban air quality and the host of associated human health impacts. The World Health Organization (WHO) estimates that, every year, around seven million deaths occur due to exposure to air pollution. Even though the solar irradiance is critical in driving atmospheric photo-chemistry via photolysis, it is marked by a severe paucity of data at the neighborhood scale.

In order to achieve the goal, the first key step is the use of multi-variate non-parametric non-linear machine learning, to accurately calibrate a set of low-cost sensors costing around USD 20 against a NIST calibrated reference instrument. The second step is physically understanding the relative importance of the various factors involved in the calibration. These factors are objectively determined by using explainable machine learning approaches. The plans and circuit diagrams for building these sensors, as well as the calibration code, are publicly available.

1.2. Solar Irradiance

The sun is a hot plasma sphere (73% hydrogen, 25% helium and 2% of heavier elements) heated to incandescence by nuclear fusion reactions in the core. The photosphere of the sun has an effective temperature of 5772 K, with an emission spectra close to that of a black body. The electromagnetic energy reaching the top of the Earth's atmosphere from the sun ranges from 100 nm to 1 mm with a peak at around 500 nm [11].

As the sunlight passes through the Earth's atmosphere, it is absorbed and scattered by various atmospheric components related to the path length through the atmosphere (Figure 1), which is a function of the solar zenith angle (Figure 2). The solar irradiance incident on the Earth's surface during the daytime is a function of the Earth's distance from the sun [15] (varies with season due to the ellipse orbit [16]), the solar zenith angle (can be calculated from latitude, longitude and the local solar time), the vertical profile of atmospheric light scatterers and absorbers and the surface reflectivity.

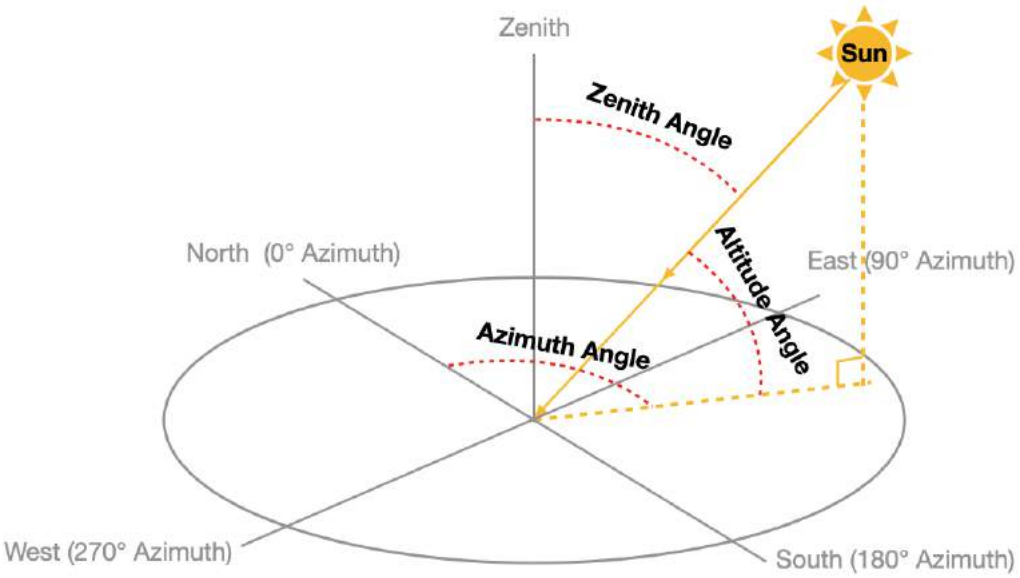


Figure 2. Schematic depicting the solar zenith angle, solar altitude angle and solar azimuth angle.

Typically, a vertical profile through the atmosphere is split up into regions based on the temperature gradient with increasing height, e.g., the troposphere, stratosphere, mesosphere and thermosphere [17]. Significant absorption by ozone of the incoming sunlight occurs in the stratosphere (10–50 km). For typical levels of stratospheric ozone, this light absorption warms the stratosphere and prevents short wave UV at $\lambda < 310$ nm (which is harmful to life) from reaching the surface of the Earth.

In a clear sky and without nearby structures such as trees or buildings, solar irradiance is primarily dependent upon some simple factors, such as the temperature of the sun, Earth's distance and solar zenith angle. However, in a real-world environment, clouds, particulate matter (PM), trees and buildings influence the intensity of solar irradiance on the ground and make it difficult to estimate the irradiance spectrum. Therefore, it is necessary to use high-resolution spectrophotometers to measure the solar spectral irradiance. However, these devices are quite expensive and cannot be widely distributed in large numbers. Thus, we proposed a machine learning method, which works with some low-cost light sensors, in order to achieve competitive performance as well as high-resolution spectrophotometers. The comparisons between the observations of the reference sensor and our machine learning calibrated low-cost sensor ensemble have been shown in Sections 4.3 and 5. With our machine learning model, we recreated the wavelength-resolved spectrum from 360 nm to 780 nm and obtained an accurate spectrum of atmospheric absorption.

2. Measurements and Data Sets

We used two types of light sensors, a NIST calibrated reference sensor (Konica Minolta CL-500A with a cost of ≈USD 6000) and an ensemble of low-cost sensors (Adafruit TSL2591, VEML6075 and AS7262, each costing just a few dollars) to collect the solar irradiance in a same environment. We further calibrated the low-cost sensors against the reference sensor by using machine learning.

The reference Minolta CL-500A provides the irradiance every nm from 360–780 nm, as well as the total illuminance (Figure 3). The codes for collecting the irradiance data from the reference sensor can be found in the following GitHub repository: https://github.com/yichigo/Minolta-Sensor, accessed 18 March 2021.

Figure 3. Konica Minolta CL-500A Illuminance Spectrophotometer measures the wavelength range of 360–780 nm.

The various low-cost sensors (Figure 4) sold by open-source hardware company Adafruit Industries are as follows.

AS7262 provides a measurement of the intensity over the broad spectral regions corresponding to "Violet" (450 nm), "Blue" (500 nm), "Green" (550 nm), "Yellow" (570 nm), "Orange" (600 nm) and "Red" (650 nm) light. The circuit design of Adafruit's AS7262 can be found in https://learn.adafruit.com/adafruit-as7262-6-channel-visible-light-sensor, accessed 22 July 2021.

TSL2591 has a sensitivity to the wavelength range 300–1000 nm, from UV to NIR. It gives the raw counts of "Visible", "IR (Infrared)" photons and the value of "Lux". The circuit design of Adafruit's TSL2591 can be found in https://learn.adafruit.com/adafruit-tsl2591, accessed 22 July 2021.

VEML6075 has a sensitivity relative to the wavelength range of 200–400 nm (UV), where the UVA (UVB) channel has a peak sensitivity at 365 nm (330 nm). It also provides "Visible Compensation" and "IR (Infrared) Compensation" for calculating the UVA (UVB) from the raw counts. In this study, we used the raw counts of UVA, UVB and the compensation values. In addition, VEML6075 provides a calculated "UV index" value, which is negatively correlated with the UV intensity. The circuit design of Adafruit's VEML6075 can be found in https://learn.adafruit.com/adafruit-veml6075-uva-uvb-uv-index-sensor, accessed 22 July 2021.

The codes for receiving the data from the low-cost sensors are in the following Github repository: https://github.com/mi3nts/UTDNodes/tree/master/firmware/nanoLightUTD, accessed 22 July 2021.

The reference and low-cost sensors were co-located in the same outdoor environment, at the University of Texas at Dallas Waterview Science and Technology Center, 7919 Waterview Parkway, Richardson, TX 75080, from December 2019 to April 2020. They made the observations every 3 s. The data were collected and saved in an NAS hard drive connected in the same network.

(a) AS7262 (b) TSL2591 (c) VEML6075

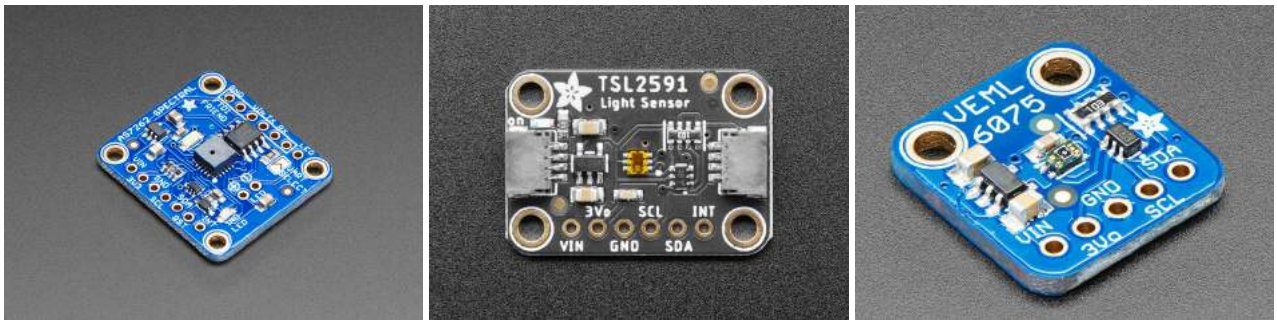


Figure 4. Low-cost light sensors: TSL2591, VEML6075 and AS7262 sold by the open-source hardware company, Adafruit Industries.

3. Machine Learning and Workflow

Machine learning was used to calibrate the inputs provided by the suite of low-cost sensors against the reference sensor. Machine learning is a subset of artificial intelligence where we "learn by example". It optimizes an empirical mathematical model by using examples rather than explicitly programming a deterministic model [18]. This technique is widely used in many areas, such as data mining [19], game strategies [20], healthcare and medical research [21–23], computer vision [24,25] and environmental science [26,27]. Machine learning can be divided into three categories: supervised learning, unsupervised learning and reinforcement learning. In this study, we used supervised learning, which trains the model by using a set of examples in a training data set that includes both input features and output targets.

3.1. Data Preprocessing

We have collected the data from different sensors. In order to merge these data into a same data set, we resampled the data at every 10 s and merged the different data sources by matching the time. After that, we dropped the NaN values and duplicated data samples. We also calculated the solar zenith angle and solar azimuth angle from the latitude, longitude and UTC time.

The preprocessed data, "MINTS Light Sensor Calibration Dataset", are publicly available on Zenodo [28].

3.2. Input Features and Output Targets

The input features are the preprocessed data given by the low-cost sensors: TSL2591, VEML6075 and AS7262. The output targets are the irradiance bins on 421 wavelengths (360–780 nm) given by the Minolta CL-500A reference sensor. The 14 input features are listed below.

Features	Unit	Min	Max	Mean	StDev
Violet (450 nm)	counts	0	2717	370.67	412.56
Blue (500 nm)	counts	0	4165	528.13	611.57
Green (550 nm)	counts	0	4619	546.68	664.90
Yellow (570 nm)	counts	0	4963	573.22	710.80
Orange (600 nm)	counts	0	3646	411.31	519.83
Red (650 nm)	counts	0	3826	425.41	545.63
IR (Infrared)	counts	0	65,535	28,500.35	27,263.43
Visible	counts	0	51,573	12,722.19	17,871.05
Lux	lux	−1	2207.41	127.99	381.92

Features	Unit	Min	Max	Mean	StDev
UVA (Raw)	counts	0	40,296	2492.98	3400.20
UVB (Raw)	counts	0	44,768	2682.46	3686.13
Visible Compensation	counts	0	12,045	782.75	1026.57
IR (Infrared) Compensation	counts	0	8596	456.58	699.96
UV Index	N/A	−1.17	1.07	−0.05	0.08

The measurement range of the low-cost sensor TSL2591 is not large enough; thus, some of the features may produce unreasonable values. For example, "IR (Infrared)" or "IR (Infrared)" + "Visible" cannot be greater than 65,535 (the maximum of 16-bit integer $2^{16} - 1$). Thus, if "IR (Infrared)" is very large, then "Visible" decreases to zero. Another feature, "Lux", should possess a zero-value minimum; however, in a very bright environment, it produces -1 rather than a large value. Therefore,"Visible" and "Lux" values may be negatively correlated with irradiance, although they should be positively correlated in physics. VEML6075 sensor produces the "UV Index", which is also negatively correlated with the irradiance, and most of the values are between -0.5 and 0.05.

The output targets given by the Minolta sensor have 421 columns, as listed below.

Targets	Unit	Min	Max	Mean	StDev
Irradiance at 360 nm	$W/m^2/nm$	0	0.069338	0.010471	0.011335
...	...	...	...	...	...
Irradiance at 780 nm	$W/m^2/nm$	0	0.337675	0.032955	0.046052

We merged the above features and targets into a single data set. It was randomly shuffled and split up into two portions, 80% of the data were used for training a suite of machine learning algorithms, the remaining 20% was used to independently test the generalization of the machine learning models.

We noticed that there is multi-collinearity between some of the features. For example, the AS7262 sensor's data "Violet", "Blue", "Green" and so on are highly correlated with each other; thus, the machine learning model may focus more on one of them by chance, and their feature importance may also be ranked by chance. Thus, we used principal component analysis (PCA) to remove the multi-collinearity.

3.3. Principal Component Analysis (PCA)

The process of PCA can be described as the following: In the N-dimension scaled feature space, we can find a direction that maximizes the variance of the data. We then use that direction as our first principal direction, and we project the data set into an N-1 dimension space by removing the first principal direction. We repeat this process for M times, where $M \leq N$, and obtain a transformed data set in the principal dimensions [29,30].

Typically, people use the PCA technique to reduce the dimensions of the data. However, in this study, we did not reduce the dimension. The input data is still 14-dimension after the PCA process. We only used PCA to remove the multicollinearity between the input features. After training the model, we ranked the feature importances in principle dimensions.

3.4. Artificial Neural Network

Artificial Neural Network (ANNs) are one of the many types of machine learning algorithms. The idea central to ANNs is to mimic the neural network found in the human brain in order to solve complex non-linear problems [31].

The left panel of Figure 5 shows an example node (neuron) in a neural network that has four inputs x_i, four weights w_i and one bias. The linear function produces $bias + \sum_{i=1}^{4} x_i w_i$ and followed by an activation function. A neuron such as this can be used as a linear classifier by optimizing the weights and bias.

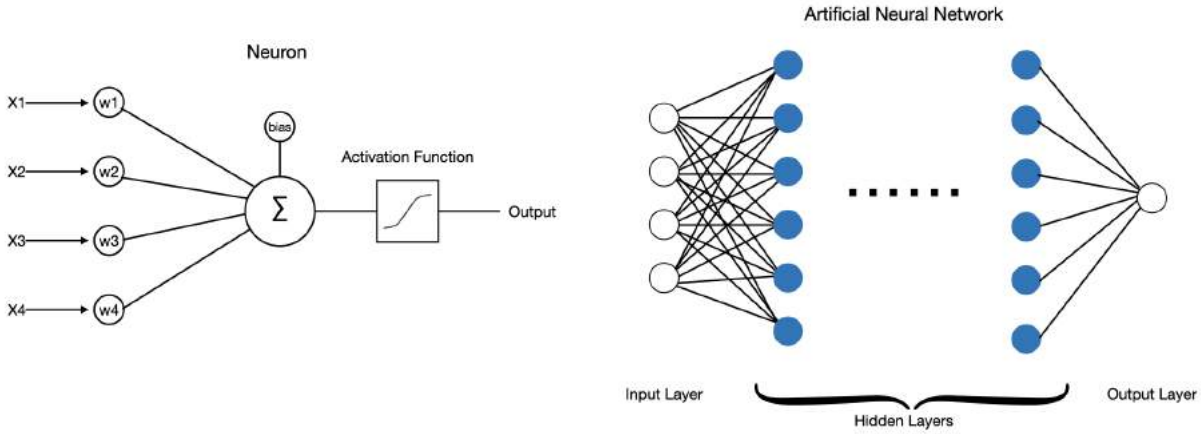


Figure 5. The left panel shows an example node in a neural network, where w_i is the weights for each input x_i, and the linear function produces $bias + \sum_i x_i w_i$ that is passed to an activation function. The right panel shows an example of artificial neural network (ANN) with single output, where the blue nodes in hidden layers and the white node in the output layer are neurons.

Artificial neural networks (ANN) are composed of an input layer, one or more hidden layers and an output layer [32]. As shown in the right panel of Figure 5, each layer has one or more neurons. The input layer receives input features, then feed into the first hidden-layer. The outputs of the first hidden-layer become the inputs to the second hidden-layer and so on. The data passes through all the hidden layers and finally arrives at the output layer. Each node in the hidden layers or output layer has an associated set of weights and bias, and it is followed by an activation function. By using back-propagation [33,34], the gradient of the loss function can be computed with respect to the weights of the network, and the weights can be optimized to fit the model on the training data.

Here, we used a multilayer neuron to calibrate the input features provided by the low-cost light sensors against the data provided by the Minolta reference sensor [35–37]. A multilayer neuron (MLP) is a class of feedforward artificial neural network (ANN). We built a three-hidden-layer MLP model, where the sizes of the hidden layers are (64-128-256), and the size of the input (output) layer is 14 (421). We used the ReLu activation function after each hidden layer. The loss function is the mean squared error, and it is optimized by the Adam optimizer. The L2 regularization penalty parameter is 10^{-5}, and we did not introduce any batch normalization [38] or dropout layer [39–41] here.

There are 345,677 date samples. We randomly shuffled the data, then used 80% of the data for training and 20% of the data for testing. In the training data set, the model is actually trained on 90% of the data, and it is valid on the other 10%. We used the standard scaler $x' = (x - \bar{x})/\sigma$ to scale the input features (before PCA and before ANN model) and output targets based on the training data set, where x can be any column of feature or target, $\bar{x}$ is the mean value of x and σ is the standard deviation. The size of mini-batches is 200 for stochastic optimizer.

We set the initial learning rate as 10^{-3} and trained the model for 40 epochs until the R^2 validation score was not improved by at least 10^{-4} for 10 consecutive epochs. Then, we divided the learning rate by 10 and repeated the process for an additional 12 epochs.

3.5. Workflow

In general, the picture of the workflow is shown in Figure 6.

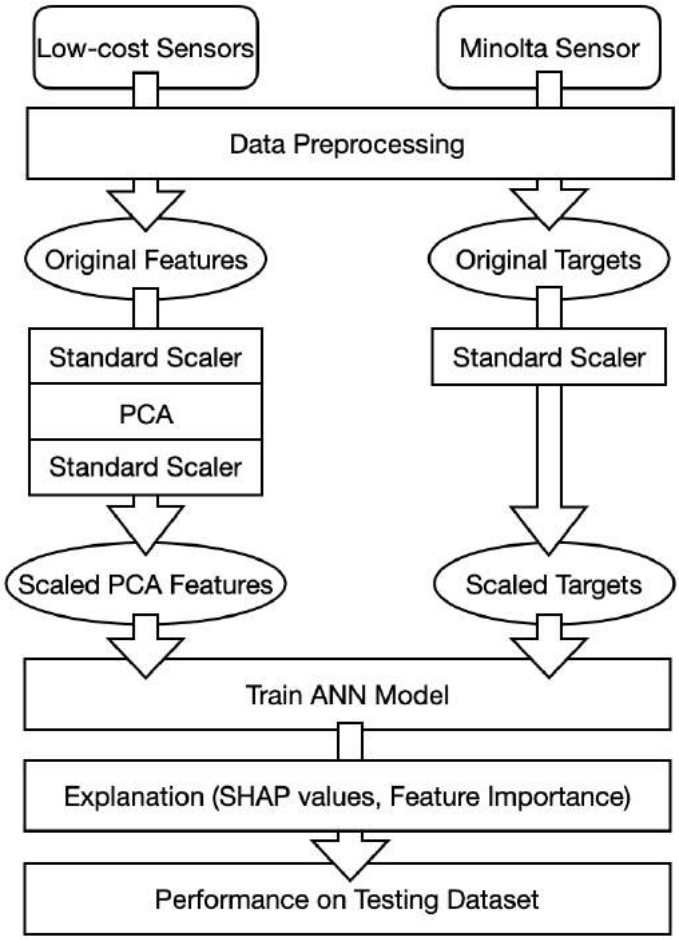


Figure 6. Workflow of the light sensor calibration.

We collected the data form the low-cost sensors and the reference Minolta sensor and performed preprocessing in order to combine and clean the data. Then, we used standard scaler and PCA techniques to generate the scaled PCA input features and scaled output targets for the ANN model. We trained the ANN model on the training data set and explained the model with SHAP values and feature importance. Finally, we tested the performance on the testing data set.

4. Machine Learning for Low-Cost Light Sensor Calibration of Wavelength Resolved Irradiance

4.1. Whole Spectrum Calibration Model (360–780 nm)

The ANN regression model was trained to calibrate and provide the entire spectrum only from the data provided by the low-cost sensor suite (i.e., to observe if we could reproduce the observations made by a USD 6000 by using sensors costing only a few dollars). For this entire spectrum calibration model, we used 421 neurons as the output layer, and one neuron for each wavelength measured by the Minolta reference sensor.

The upper left panel of Figure 7 shows the scatter diagram showing the performance of this model by comparing the estimated value (y-axis) and the actual value (x-axis) of the irradiance, from 360 nm to 780 nm, observed by the reference sensor. The coefficient of determination (R^2) is 0.9987 on the training data and is 0.9983 on the testing data.

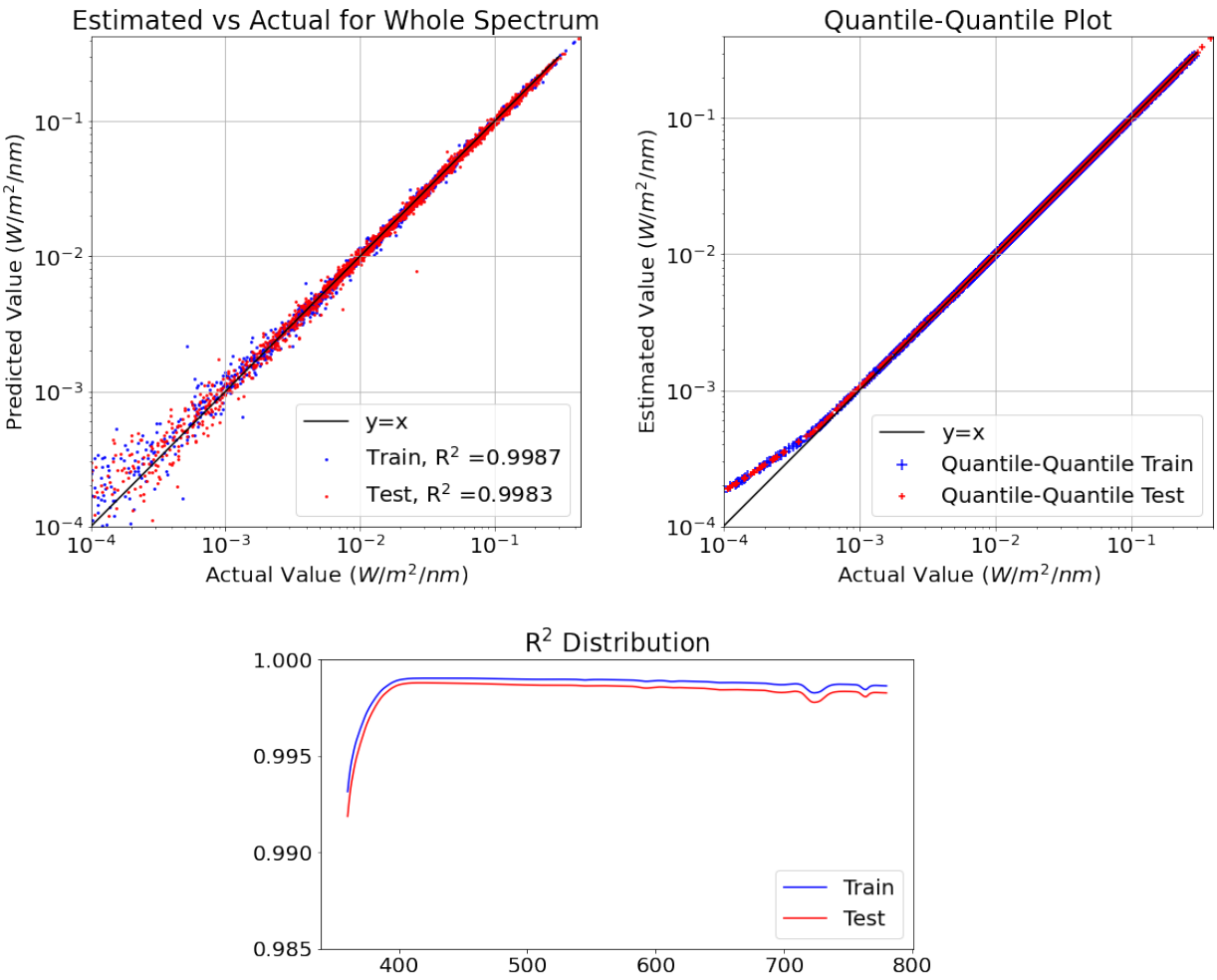


Figure 7. Whole spectrum model performance and the R2 scores on different wavelengths.

The upper right panel shows the quantile–quantile plot, which compares the shape of the probability distribution of our estimates against the shape of the probability distribution of the actual observations. We can observe the distributions of the actual observations, and our estimated values are almost the same above 10^{-3} and slightly different below that.

The lower panel shows the R^2 values for each wavelength between 360 and 780 nm; we observe that that all the coefficient of determination on the testing data are above 0.99. The machine learning model has performed well at all wavelengths.

4.2. The Relative Importance of the Machine Learning Inputs

It is very helpful to understand the relative importance of the machine learning inputs. Let us take a look at a couple of approaches that estimate the relative importance of the machine learning inputs in performing the calibration of the low-cost sensors.

4.2.1. Shapley Value: An Explainer of Machine Learning Models

The Shapley (SHAP) value was introduced by Lloyd Shapley in 1951 [42]. It is a game-theoretic approach for calculating the marginal contribution of each player in a cooperative game. In machine learning, we calculated the SHAP value for each data sample on each feature. The SHAP value indicates how much the feature value of a sample raises or decreases the target value.

The processes for calculating the SHAP value are described as follows.

Assume the value function of a teamwork is v(S), where v is an arbitrary value function which can be a math function or a machine learning model and S is a subset of the players

(features) who attended the game, which may contain x_0, x_1, x_2, ... For a given data sample, in order to calculate the value function $v(S)$, only the attended features in S use the sample values, while other features use their mean values. The contribution of the player (feature) x_i is as follows:

$$\phi_i = \sum_{S \in \{x_1, x_2, \ldots, x_N\} \setminus \{x_i\}} \frac{|S|!(N - |S| - 1)!}{N!} \Big(v(S \cup \{x_i\}) - v(S) \Big) \tag{1}$$

where $|S|$ is the number of the players (features) in S and N is the total number of the players (features). The term $v(S \cup \{x_i\}) - v(S))$ provides the marginal contribution of x_i when it joins the game in addition to S. The weight $\frac{|S|!(N-|S|-1)!}{N!}$ can be calculated from the permutation of players (features) in which $|S|!$ is the permutation of the players (features) who attended the game before x_i, $(N - |S| - 1)!$ is the permutation of the players who do not attend the game and $N!$ is the permutation of all the players (features). We needed to calculate all the cases of S that did not include x_i and summed up the weighted marginal contribution of x_i. Then, we obtained the contribution of x_i, and that is its SHAP value.

In this study, the inputs and outputs of the ANN machine learning model are scaled by $x' = (x - \bar{x})/\sigma$, where x can be any column in the features or targets, $\bar{x}$ is the mean value of x and σ is the corresponding standard deviation. The SHAP values we calculated here are the contributions to the 421 scaled targets (360–780 nm), and we take the averaged SHAP values over these targets.

In order to explain the ANN model, we used the SHAP value to show how each feature contributes to the model's output, and how each ranks the corresponding feature importance.

We plotted the SHAP values of a random subset of the data points in the following process: normalize the feature values in a color scheme, list different features in the vertical direction and list the SHAP values of each feature in the horizontal direction.

The SHAP values in the left panel of Figure 8 shows how the PCA input features impact the ANN model's output. For example, the red points on the right side means that if a larger value was the input for this feature, then the output of the model would also increase. We ranked the PCA SHAP values (left panel of Figure 8) by calculating the mean absolute value of the SHAP value. The 0th principle component contributes the most to the models outputs, and other components also contribute a little.

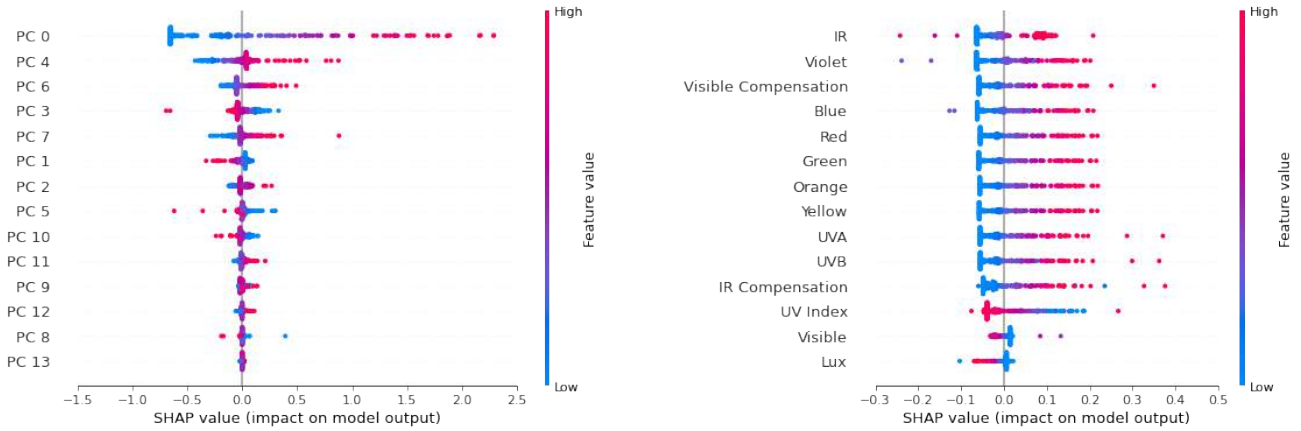


Figure 8. SHAP value of our MLP model, which calibrates the whole spectrum, shows how the PCA features impact the models' output. The red (blue) color denotes the high(low) value of each feature and the position of point on the x direction shows the impact on the target. The left panel shows the principle components' SHAP values, and the right panel shows the original features' SHAP values from the 0th principle component by using linear approximation.

We are able to linearly transform the 0th principle component's SHAP values to the original features with the first order contribution ratio $\frac{a_i(x_i-\bar{x}_i)}{\sum_i a_i(x_i-\bar{x}_i)}$, where x_i is the i-th original feature, $\bar{x}_i$ is the mean value of x_i and a_i is the coefficient of PCA transformation, as shown in the right panel of Figure 8. The original features, namely the red, orange, green, yellow, blue, UVA, UVB and IR (infrared) positively impact the model's output, while some other features such as the UV index, Visible and Lux features negatively impact the model's output. Please notice that we only simply use this linear approximation to visualize the main part (from 0th principle component) of the contributions of the original features. However, we should not sum up these original features' SHAP values by combining all the principle components because the original features have strong multi-collinearity.

4.2.2. Feature Importance

For machine learning models such as ANN, we can calculate feature importance from the mean absolute value of the Shapley (SHAP) values for each feature.

In our MLP model for calibrating the whole spectrum, we calculate the feature importance from the SHAP values, as shown in Figure 9. We used the red (blue) color to indicate a feature that positively (negatively) impacts the model. The positive/negative impact means that if a feature's value increases with the other features remaining unchanged, then the output value of the model is more likely to increase/decrease. For example, in the Figure 8 of SHAP values, the 0th principle component has a positive impact on the model's output since the red (blue) points, with high (low) feature values, are on the right (left) side, which raises (decreases) the model's output. In most of the models, the sign of positive and negative can also be easily calculated from the sign of the correlation coefficient between the feature and the target. The feature importances are ranked in descending order, and we are able to figure out which features are important for predicting our output targets. We note that most of the variables provided by the low-cost light sensor suite provide useful information.

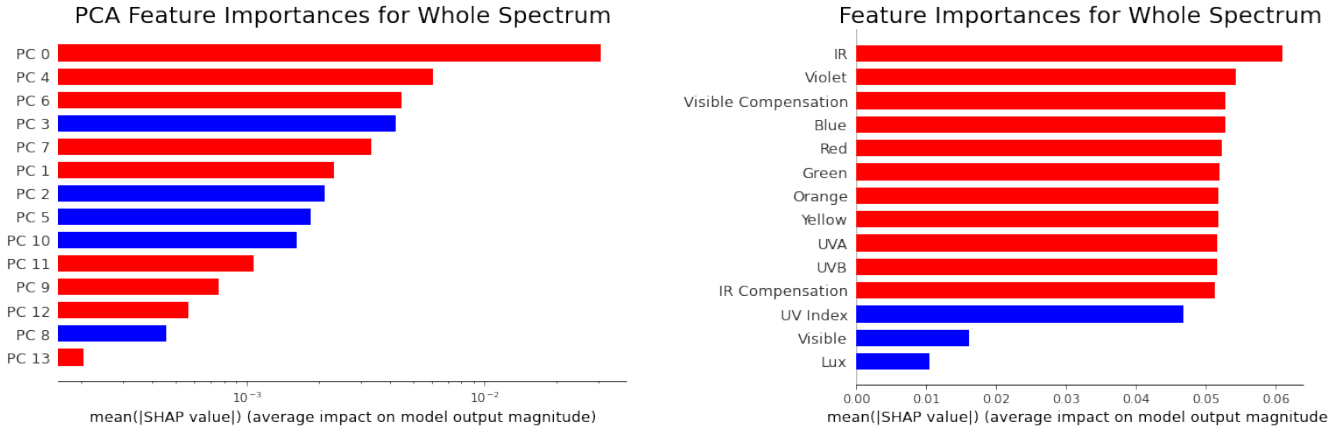


Figure 9. (**Left**) PCA feature importance of our MLP model for whole spectrum calibration; we used a log_{10} scale in the x direction. (**Right**) The original feature importances from the linearly splitting of the 0th principle component's SHAP values. A feature with red (blue) color has positive (negative) impact on the model's output.

Now, we have observed that machine learning can provide an effective calibration of the low-cost sensors, and we are familiar with the relative importance of the input parameters; let us use this calibration to examine the temporal variability that was measured.

4.3. Applying the Calibration to Provide an Irradiance Spectrum

We picked at random some time periods from the testing data set and used our neural network model to provide the full spectral irradiance from 360 nm to 780 nm by only using the data provided by the low-cost USD 20 sensor ensemble (Figure 10). We estimated the

full spectral irradiance (middle panel) from the low-cost sensor ensemble and compared this with the observed value from the reference NIST calibrated sensor (upper panel). Our ANN model successfully reproduced the high resolution spectrum by only using the data from the low-cost sensors.

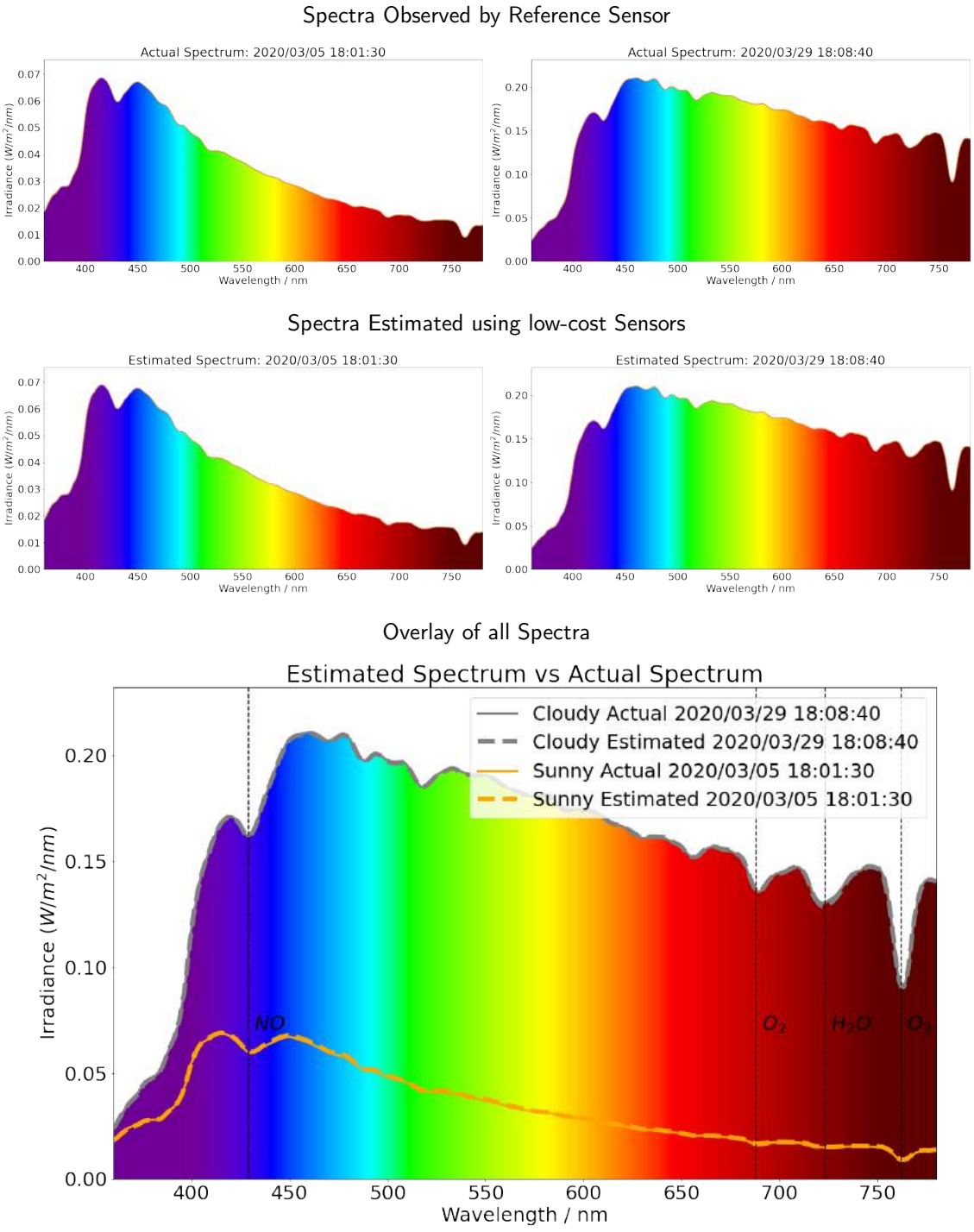


Figure 10. The full spectrum intensity (middle panel) predicted using the data from the low-cost sensors, and the full spectrum ANN compared with the spectra observed by the reference sensor (upper panel). The daily spectra over the entire day have been shown in Section 5. We can clearly observe the role of the variable weather conditions, such as changing cloudiness. It was sunny on 5 March 2020, and this can be observed by noting the higher spectral irradiance at the shorter wavelengths, indicating a blue sky day. In comparison, it was cloudy on 29 March 2020. Note how the spectral irradiances from the blue to red portion of the spectrum are much more similar than on March 5; the sky was closer to white than blue. Our ANN performed well for both clear sky and cloudy conditions, and the data from the low-cost sensors could be used to effectively reproduce the full spectra, including the atmospheric absorption bands.

We can clearly observe the role of the variable weather conditions, such as changing cloudiness. When the direct normal irradiation (DNI) is stopped by the surrounding buildings, the photons reflected by the clouds increase the total intensity of irradiance. Furthermore, as the light scattering is wavelength dependent, the shape of the spectrum changes on a cloudy day. It was sunny on 5 March 2020. Note the higher spectral irradiance at the shorter wavelengths, indicating a blue sky day. By comparison, it was cloudy on 29 March 2020. Note how the spectral irradiances from the blue to red portion of the spectrum are much more similar than on March 5; the sky was closer to white than blue. Our ANN performed well for both clear sky and cloudy conditions, and the data from the low-cost sensors could be used to effectively reproduce the full spectra.

In the bottom panel of Figure 10, we overlay the actual solar spectrum and estimated solar spectrum for both sunny (5 March 2020) and cloudy (29 March 2020) cases. Furthermore, the figure displays the atmospheric absorption spectrum, including nitric oxide (NO) at 429 nm, oxygen (O_2) at 688 nm and 762 nm and water vapor (H_2O) from 720 nm to 730 nm [43]. Using data from the low-cost sensors, our machine learning model correctly obtained all of these absorption bands and showed the potential for detecting changes of atmospheric components.

Our ANN model was trained on the MINTS light sensor calibration data set [28], the codes for training the model and generating all the figures can be found in the following Github repository [44]: https://github.com/yichigo/Light-Sensors-Calibration, accessed 25 August 2021.

5. The Observed Diurnal Variation in Wavelength Resolved Irradiance

Figure 11 shows the observations used in the testing data set for the entire three month period. As expected, the solar zenith is a key factor in determining solar irradiance. Secondly, we note that, on cloudy days, the multiple scattering of light from airborne aerosols and clouds coupled with the surface reflection of the scattered light increases the surface irradiance substantially (Figure 11). In Figure 11, we observe that the surface irradiance observed by the reference sensor and that is estimated by the low-cost ensemble of sensors calibrated using machine learning agree very well; the blue and red points overlay one another so precisely that they produce the appearance of magenta points. In each panel, the solid curves close to the bottom shows the averaged irradiance for the sunny days, and the cloudy days produce a higher value of irradiance due to the trapping of photons by multiple scattering from the clouds and surface reflection of the scattered light.

Figure 12 compares the wavelength (y-axis) and UTC time (x-axis) resolved daily spectra for a sunny day on the left, and for a cloudy day on the right. In both cases, we can observe the key role that both the solar zenith angle and the cloudiness plays in determining the intensity of sunlight at the surface of the Earth.

We can observe that the surface irradiance wavelength distribution is a strong function of the conditions and that atmospheric multiple scattering of sun light plays a substantial role. For a sunny day, the solar irradiance spectrum peaks in the violet and blue part of the spectrum and is close to that of a black body at around 5772 K. On a cloudy day the multiple light scattering and surface reflection traps photons, thereby enhancing the intensity at the longer wavelengths. The cloud water drops result in Mie scattering of the sunlight. Thus, the diffuse horizontal irradiance (DHI) is actually greater on a cloudy day than it is on a sunny day. This is very evident when we take a snapshot at an instant in time and examine the shape of the spectrum as a function of wavelength and compare a sunny and cloudy day (Figure 13).

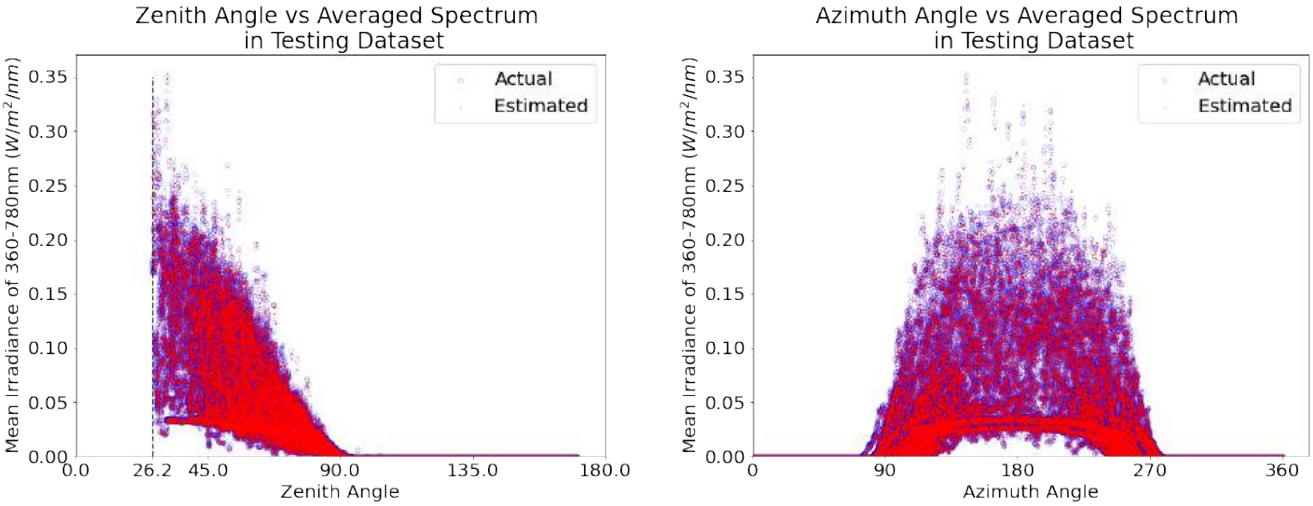


Figure 11. The mean irradiance from 360–780 nm as a function of solar zenith angle (**left panel**) and solar azimuth angle (**right panel**) from the testing data set for the entire three month period. The blue circles show the actual irradiance; the overlaid red points show the machine learning estimates. We observe that the surface irradiance observed by the reference sensor and that is estimated by the low-cost ensemble of sensors calibrated using machine learning agree very well; the blue and red points overlay one another so precisely that they produce the appearance of magenta points. In each panel, the solid curves close to the bottom shows the averaged irradiance for the sunny days, and the cloudy days produce higher values of irradiance due to the trapping of photons by multiple scattering from the clouds and surface reflection of the scattered light.

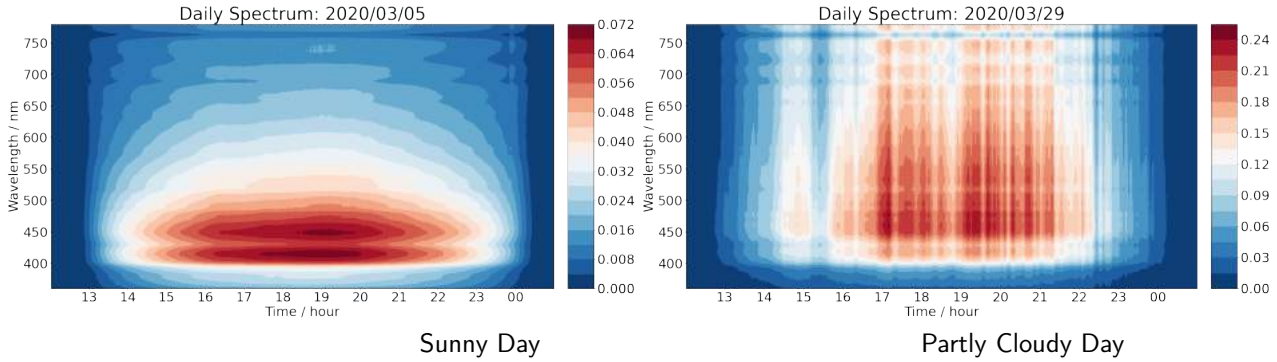


Figure 12. A comparison of the wavelength (*y*-axis) and UTC time (*x*-axis) resolved daily spectra for a sunny day on the left and for a cloudy day on the right. The color denotes the solar irradiance in unit $W/m^2/nm$. Both spectra were collected using the Minolta CL-500A Illuminance Spectrophotometer.

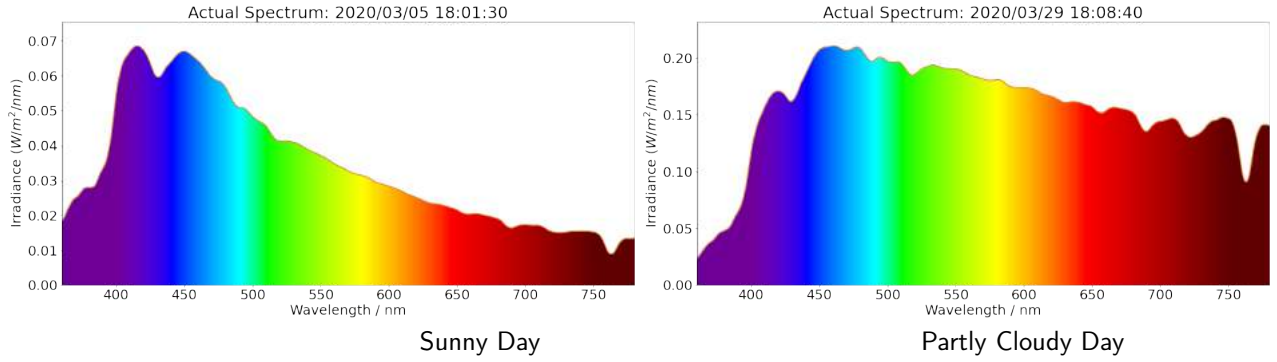


Figure 13. Spectra of solar irradiance measured by Minolta CL-500A Illuminance Spectrophotometer on a certain time, where x value is the wavelength, y value is the intensity of irradiance in unit $W/m^2/nm$ and the color denotes the visible color of the corresponding wavelength.

The sunny day irradiance spectrum is close to that of a black body at around 5772 K. The partly cloudy spectrum has enhanced intensity, particularly at longer wavelengths due to Mie scattering in the clouds. The Mie scattering on the clouds generates greater diffuse horizontal irradiance (DHI) and makes the sky brighter.

6. Conclusions

A neural network algorithm was able to effectively calibrate an ensemble of low-cost light sensors and generate a high resolution wavelength resolved solar irradiance spectrum. The roles of the solar illumination geometry (such as the solar zenith angle) and the weather conditions (such as the cloudiness) were clearly evident. These low-cost light sensor packages are currently being deployed across cities in the Dallas Fort Worth (DFW) area for characterizing both the temporal and spatial scales of solar irradiance. All the sensor circuit designs and calibration codes have been made open source.

Author Contributions: Conceptualization, D.J.L.; Data curation, Y.Z. and L.O.H.W.; Formal analysis, Y.Z.; Funding acquisition, D.J.L.; Investigation, Y.Z. and D.J.L.; Methodology, Y.Z. and D.J.L.; Project administration, Y.Z. and D.J.L.; Resources, L.O.H.W. and D.J.L.; Software, Y.Z. and S.T.; Supervision, D.J.L.; Validation, D.J.L.; Visualization, Y.Z.; Writing—original draft, Y.Z.; Writing—review & editing, Y.Z. and D.J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by USAMRMC Award Number W81XWH-18-1-0400. The authors acknowledge the Texas Research and Education Cyberinfrastructure Services (TRECIS) Center, NSF Award #2019135, and the University of Texas at Dallas for providing HPC, visualization, database or grid resources and support that have contributed to the research results reported within this paper. URL: https://trecis.cyberinfrastructure.org/, accessed 22 July 2021.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Christopher Simmons is gratefully acknowledged for his computational support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chandrasekhar, S. *Radiative Transfer*; Dover Publications: Mineola, NY, USA, 1960.
2. Lenoble, J. *Radiative Transfer in Scattering and Absorbing Atmospheres: Standard Computational Procedures*; A. DEEPAK Publishing: Hampton, VA, USA, 1985.
3. Lary, D.J.; Pyle, J.A. Diffuse radiation, twilight, and photochemistry—I. *J. Atmos. Chem.* **1991**, *13*, 373–392. [CrossRef]
4. Lary, D.J.; Pyle, J.A. Diffuse radiation, twilight, and photochemistry—II. *J. Atmos. Chem.* **1991**, *13*, 393–406. [CrossRef]
5. Deutschmann, T.; Beirle, S.; Friess, U.; Grzegorski, M.; Kern, C.; Kritten, L.; Platt, U.; Prados-Roman, C.; Pukīte, J.; Wagner, T.R.; et al. The Monte Carlo atmospheric radiative transfer model McArtim: Introduction and validation of Jacobians and 3D features. *J. Quant. Spectrosc. Radiat. Transf.* **2011**, *112*, 1119–1137. [CrossRef]
6. Hartmann, D.L. (Ed.) Atmospheric Radiative Transfer and Climate. In *International Geophysics*; Elsevier: Amsterdam, The Netherlands, 2016.
7. Buehler, S.; Mendrok, J.; Eriksson, P.; Perrin, A.; Larsson, R.; Lemke, O. ARTS, the Atmospheric Radiative Transfer Simulator—Version 2.2, the planetary toolbox edition. *Geosci. Model Dev.* **2017**, *11*, 1537–1556. [CrossRef]
8. Zhang, F.; Shi, Y.; Wu, K.; Li, J.; Li, W. Atmospheric Radiative Transfer Parameterizations. In *Understanding of Atmospheric Systems with Efficient Numerical Methods for Observation and Prediction*; IntechOpen: London, UK, 2019.
9. Gordon, I.E.; Babikov, Y.L.; Barbe, A.; Benner, D.C.; Bernath, P.F.; Birk, M.; Bizzocchi, L.; Boudon, V.; Brown, L.R.; Chance, K.; et al. The HITRAN2012 molecular spectroscopic database. *J. Quant. Spectrosc. Radiat. Transf.* **2013**, *130*, 4–50.
10. Noelle, A.; Hartmann, G.; Fahr, A.; Lary, D.; Lee, Y.P.; Limão-Vieira, P.; Locht, R.; Martín-Torres, F.J.; McNeill, K.; Orlando, J.; et al. *UV/Vis+ Spectra Data Base (UV/Vis+ Photochemistry Database)*, 12th ed.; Science-softCon: Maintal, Germany, 2019; ISBN 978-3-00-063188-7.
11. Brasseur, G.; Solomon, S. *Aeronomy of the Middle Atmosphere*, 2nd ed.; D.Reidel Publishing Company: Dordrecht, The Netherlands, 1986.

12. Shanmugam, V.; Shanmugam, P.; He, X. New algorithm for computation of the Rayleigh-scattering radiance for remote sensing of water color from space. *Opt. Exp.* **2019**, *27*, 30116–30139. [CrossRef]
13. Krishnan, R.S. The scattering of light by particles suspended in a medium of higher refractive index. *Proc. Indian Acad. Sci.—Sect. A* **1934**, *1*, 147–155. [CrossRef]
14. Laeng, B.; Sirois, S.; Gredebäck, G. Pupillometry: A Window to the Preconscious? *Perspect. Psychol. Sci. J. Assoc. Psychol. Sci.* **2012**, *7*, 18–27. [CrossRef]
15. Boxwell, M. *Solar Electricity Handbook: A Simple, Practical Guide to Solar Energy: How to Design and Install Photovoltaic Solar Electric Systems*; Greenstream Publishing: Coventry, UK, 2012.
16. Wayne, R.P. *Chemistry of Atmospheres*, 3rd ed.; Oxford University Press: Oxford, UK, 2000.
17. Brasseur, G.P.; Orlando, J.J.; Tyndall, G.S. *Atmospheric Chemistry and Global Change*; Oxford University Press: Oxford, UK, 1999.
18. Koza, J.R.; Bennett, F.H.; Andre, D.; Keane, M.A. Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In *Artificial Intelligence in Design '96*; Springer: Cham, The Netherlands, 1996; pp. 151–170. [CrossRef]
19. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From data mining to knowledge discovery in databases. *AI Mag.* **1996**, *17*, 37–37.
20. Samuel, A.L. Some studies in machine learning using the game of checkers. II—Recent progress. In *Comput. Games I*; Springer: Berlin/Heidelberg, Germany, 1988; pp. 366–400.
21. Dudoit, S.; Fridlyand, J.; Speed, T.P. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **2002**, *97*, 77–87. [CrossRef]
22. Yarkoni, T.; Poldrack, R.A.; Nichols, T.E.; Van Essen, D.C.; Wager, T.D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **2011**, *8*, 665. [CrossRef]
23. Pereira, F.; Mitchell, T.; Botvinick, M. Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage* **2009**, *45*, S199–S209. [CrossRef]
24. Bhavsar, P.; Safro, I.; Bouaynaya, N.; Polikar, R.; Dera, D. Machine learning in transportation data analytics. In *Data Analytics for Intelligent Transportation Systems*; Elsevier: Amsterdam, The Netherlands, 2017; pp. 283–307.
25. Hagenauer, J.; Helbich, M. A comparative study of machine learning classifiers for modeling travel mode choice. *Exp. Syst. Appl.* **2017**, *78*, 273–282. [CrossRef]
26. Lary, D.J.; Alavi, A.H.; Gandomi, A.H.; Walker, A.L. Machine learning in geosciences and remote sensing. *Geosci. Front.* **2016**, *7*, 3–10. [CrossRef]
27. Huang, C.; Davis, L.; Townshend, J. An assessment of support vector machines for land cover classification. *Int. J. Remote Sens.* **2002**, *23*, 725–749. [CrossRef]
28. Zhang, Y. *MINTS Light Sensor Calibration Dataset*; Zenodo: Genève, Switzerland, 2021. [CrossRef]
29. Rao, C.R. The use and interpretation of principal component analysis in applied research. *Sankhyā Indian J. Statis.* **1964**, *12*, 329–358.
30. Pearson, K.L., III. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572. [CrossRef]
31. Haykin, S.S.; others. *Neural Networks and Learning Machines/Simon Haykin*; Prentice Hall: New York, NY, USA, 2009.
32. Liakos, K.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine learning in agriculture: A review. *Sensors* **2018**, *18*, 2674. [CrossRef] [PubMed]
33. Hecht-Nielsen, R. Theory of the backpropagation neural network. In Proceedings of the International 1989 Joint Conference on Neural Networks, Washington, DC, USA, 18–22 June 1989; Volume 1, pp. 593–605.
34. Widrow, B.; Lehr, M.A. 30 years of adaptive neural networks: Perceptron, Madaline, and backpropagation. *Proc. IEEE* **1990**, *78*, 1415–1442 [CrossRef]
35. Feng, X.; Li, Q.; Zhu, Y.; Hou, J.; Jin, L.; Wang, J. Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. *Atmos. Environ.* **2015**, *107*, 118–128. [CrossRef]
36. Wijeratne, L.O.; Kiv, D.R.; Aker, A.R.; Talebi, S.; Lary, D.J. Using machine learning for the calibration of airborne particulate sensors. *Sensors* **2020**, *20*, 99. [CrossRef] [PubMed]
37. Liang, X.; Liu, Q.M. Applying Deep Learning to Clear-Sky Radiance Simulation for VIIRS with Community Radiative Transfer Model—Part 2: Model Architecture and Assessment. *Remote Sens.* **2020**, *12*, 3825. [CrossRef]
38. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015.
39. Wan, L.; Zeiler, M.D.; Zhang, S.; LeCun, Y.; Fergus, R. Regularization of Neural Networks using DropConnect. In Proceedings of the International Conference on Machine Learning ICML, Atlanta, GA, USA, 16–21 June 2013.
40. Hinton, G.E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv* **2012**, arXiv: 1207.0580.
41. Srivastava, N.; Hinton, G.E.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
42. Shapley, L.S. *Notes on the n-Person Game—II: The Value of an n-Person Game*; RAND Corporation: Santa Monica, CA, USA, 1951.

43. Miyauchi, M. Properties of Diffuse Solar Radiation under Overcast Skies with Stratified Cloud. *J. Meteorol. Soc. Jpn. Ser. II* **1985**, *63*, 1083–1095. [CrossRef]
44. yichigo. *yichigo/Light-Sensors-Calibration: MINTSLightSensorsCalibration*; Zenodo: Genève, Switzerland, 2021. [CrossRef]

sensors

MDPI

Article

Data-Driven EEG Band Discovery with Decision Trees

Shawhin Talebi *, John Waczak, Bharana A. Fernando, Arjun Sridhar and David J. Lary

Hanson Center for Space Sciences, University of Texas at Dallas, Richardson, TX 75080, USA; john.waczak@utdallas.edu (J.W.); ashen.fernando@utdallas.edu (B.A.F.); arjun.sridhar@utdallas.edu (A.S.); david.lary@utdallas.edu (D.J.L.)

* Correspondence: shawhin.talebi@utdallas.edu

Abstract: Electroencephalography (EEG) is a brain imaging technique in which electrodes are placed on the scalp. EEG signals are commonly decomposed into frequency bands called delta, theta, alpha, and beta. While these bands have been shown to be useful for characterizing various brain states, their utility as a one-size-fits-all analysis tool remains unclear. The goal of this work is to outline an objective strategy for discovering optimal EEG bands based on signal power spectra. A two-step data-driven methodology is presented for objectively determining the best EEG bands for a given dataset. First, a decision tree is used to estimate the optimal frequency band boundaries for reproducing the signal's power spectrum for a predetermined number of bands. The optimal number of bands is then determined using an Akaike Information Criterion (AIC)-inspired quality score that balances goodness-of-fit with a small band count. This data-driven approach led to better characterization of the underlying power spectrum by identifying bands that outperformed the more commonly used band boundaries by a factor of two. Additionally, key spectral components were isolated in dedicated frequency bands. The proposed method provides a fully automated and flexible approach to capturing key signal components and possibly discovering new indices of brain activity.

Keywords: electroencephalography (EEG); EEG bands; decision tree; machine learning

Citation: Talebi, S.; Waczak, J.; Fernando, B.A.; Sridhar, A.; Lary, D.J. Data-Driven EEG Band Discovery with Decision Trees. *Sensors* **2022**, *22*, 3048. https://doi.org/10.3390/s22083048

Academic Editor: Andrea Facchinetti

Received: 7 March 2022
Accepted: 12 April 2022
Published: 15 April 2022

1. Introduction

The electrical activity produced by the brain was discovered by Richard Caton. Hans Berger later demonstrated that this activity could be recorded directly from the scalp [1]. This technique for measuring brain activity is called electroencephalography (EEG). It consists of an array of electrodes placed on the scalp that record fluctuations in electric potential arising from the activity of synchronized neural populations [2,3].

A popular method of analyzing EEG is spectral analysis. This consists of decomposing signals onto a frequency basis (Figure 1) and grouping frequencies into spectral bands (i.e., frequency ranges). A popular method for EEG signal decomposition is Welch's method which estimates a signal's spectral power density across a range of frequencies [4]. Commonly used spectral bands are: delta, theta, alpha, and beta [5].

EEG bands correspond to brain phenomena in specific brain areas and contexts. For example, alpha activity from occipital regions (i.e., visual cortex) in relaxed, awake animals track with eye closures [6]. During sleep, alpha-band activity is observed at sleep onset, also called sleep spindles (7–14 Hz), and delta waves (1–4 Hz) appear in deep sleep stages [6]. Additionally, EEG bands have been used in a variety of contexts such as: measuring cognitive load [7–9], disease diagnosis [10–12], and predicting emotions [13–15].

Despite the widespread use of established spectral bands (e.g., delta, theta, alpha, and beta), there are two potential concerns with the current standard. First, there is significant variability in band boundaries across studies, as shown in Figure 2. This disagreement may be a result of a variety of factors such as hardware, filtering, and experimental task [12]. Second, ideal band definitions may depend on individual characteristics such as age, genetics, personality, and task performance [16].

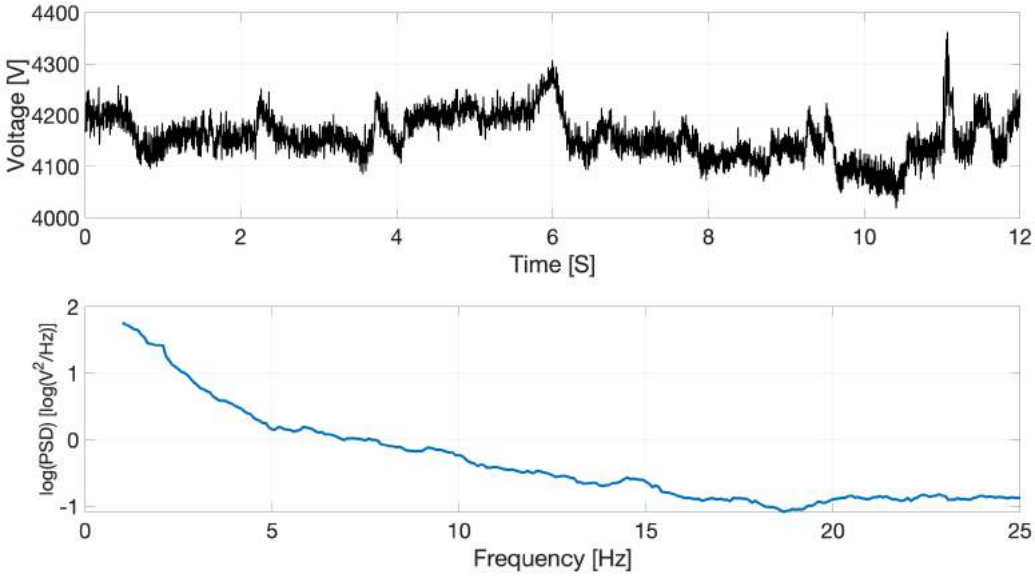


Figure 1. (**Top**) Example EEG time series signal sampled at 500 Hz. (**Bottom**) EEG signal's corresponding power spectrum, where the natural logarithm of the signal's power spectral density (PSD) is plotted against frequency.

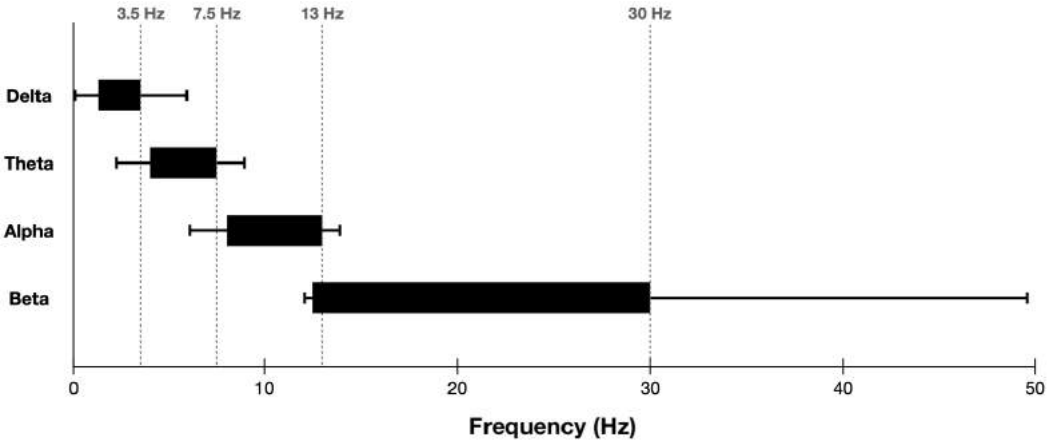


Figure 2. Box plot illustrating variability between delta, theta, alpha, and beta band boundaries across studies. Boxes indicate the typical frequency range of each band. Whiskers represent the smallest and largest band edges observed across studies. Plot adapted from figure in [12].

These concerns motivate the use of data-driven approaches for the discovery of optimal EEG band boundaries. Such an approach tailors EEG bands to a specific experimental context in an automated way. Many methodologies have been proposed to achieve this goal [16–20]. These approaches typically make use of a target variable to ground the optimization of band boundaries [17–20]. For example, learning the best choice of boundaries for classifying Alzheimer's disease [17]. A more recent approach proposed by Cohen makes use of a generalized eigendecomposition of the covariance matrix for multi-channel EEG data [16].

Here, we present a new method of EEG band discovery, that makes use of decision trees, a popular machine learning framework. Optimal bands are inferred for an input EEG power spectrum in a self-supervised way. Two key points distinguish this method from past approaches.

- Band discovery is completely self-supervised in the sense that only EEG data is used
- As the method only uses a power spectrum, it is agnostic as to how the data is generated, so it can handle both single- and multi-channel data in a variety of contexts.

2. Methods

2.1. Method Overview

An overview of the data-driven method for EEG band discovery is illustrated in Figure 3. The first step is to obtain single- or multi-channel EEG recordings. Second, a power spectrum is computed using, for example, Welch's method [4]. It is noted that information from multiple EEG channels can be aggregated into a single power spectrum in this step. Third, a set of band boundaries is derived for every possible choice of band count. Possible values range from 2 bands up to the total number of unique frequency values in the power spectrum. Fourth, an AIC-inspired quality score is computed for each choice of band count. Finally, the band boundaries with the smallest quality score are selected as the best choices.

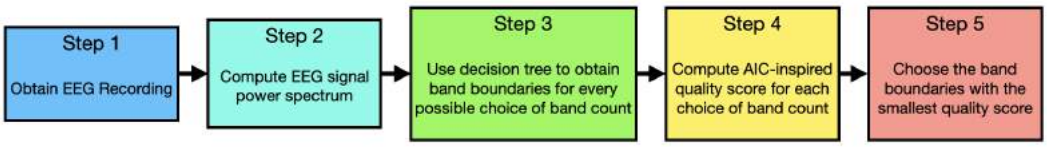


Figure 3. Conceptual overview of the data-driven methodology used in this paper.

There are two key components of this data-driven method. The first component is the use of a decision tree to obtain optimal EEG band boundaries for a specified band count. There are two main benefits to using a decision tree in this context. First, due to the structure of decision tree regression, frequency values are grouped into true bins. In other words, frequency values in a discovered band are adjacent, which may not be guaranteed by other regression techniques. The second is the ease of use. There are many efficient and ready-to-use implementations of decision tree optimization across many computational frameworks [21–24].

The second key component is an Akaike Information Criterion (AIC)-inspired quality score which serves as an objective from which the choice of band count can be optimized. As a result, this objective eliminates the need for manual entry of a band count.

2.2. Decision Trees

Decision trees are a widely-used and intuitive machine learning approach. Typically, they are used to solve prediction problems. That is, identifying a discrete target class (classification) or estimating a continuous target value (regression) from a set of predictor variables [25].

Data can be used to *grow* decision trees in an optimization process called training. Training requires a training dataset, which consists of predictor variables labeled with target values. A standard strategy for training a decision tree is recursively partitioning data via a greedy search method. The search determines the gain from each splitting option and then chooses the one that provides the greatest gain [25,26]. Splitting options are the observed predictor variable values in the training dataset. Gain is determined by the split criterion e.g., Gini impurity or mean squared error (MSE).

For example, in a regression task, data records are recursively split into two groups such that the weighted average MSE of the target value is minimized from the resulting groups. *MSE* is defined as follows.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

where, N is the total number of observations in a given partition. Y_i is the true target value for the ith frequency value. $\hat{Y}_i$ is the tree estimated target value for the ith frequency value.This splitting procedure can continue until all data partitions are pure, meaning every data record in a given partition corresponds to a single target value. Although this implies decision trees can be perfect estimators, such an approach would result in overfitting.

Therefore, the trained decision tree would not perform well on data sufficiently different than the training dataset.

One way to combat the overfitting problem is hyperparameter tuning. Hyperparameters are values that constrain the growth of a decision tree. Common decision tree hyperparameters are the maximum number of splits, minimum leaf size, and the number of splitting variables. The key result of setting decision tree hyperparameters is to limit the tree's size, which can help avoid predictions only suitable to the training dataset. In this work, we use decision tree hyperparameters to control the number of discovered frequency bands.

2.3. Band Discovery with Decision Trees

Optimal EEG frequency bands can be estimated using the decision tree framework. Here, *optimal* means the frequency groupings that best reproduce an input signal's log spectral density for a set number of bands. To achieve this goal, a decision tree is used to solve a regression problem in the usual way. A visual overview of the decision tree training in this context is shown in Figure 4.

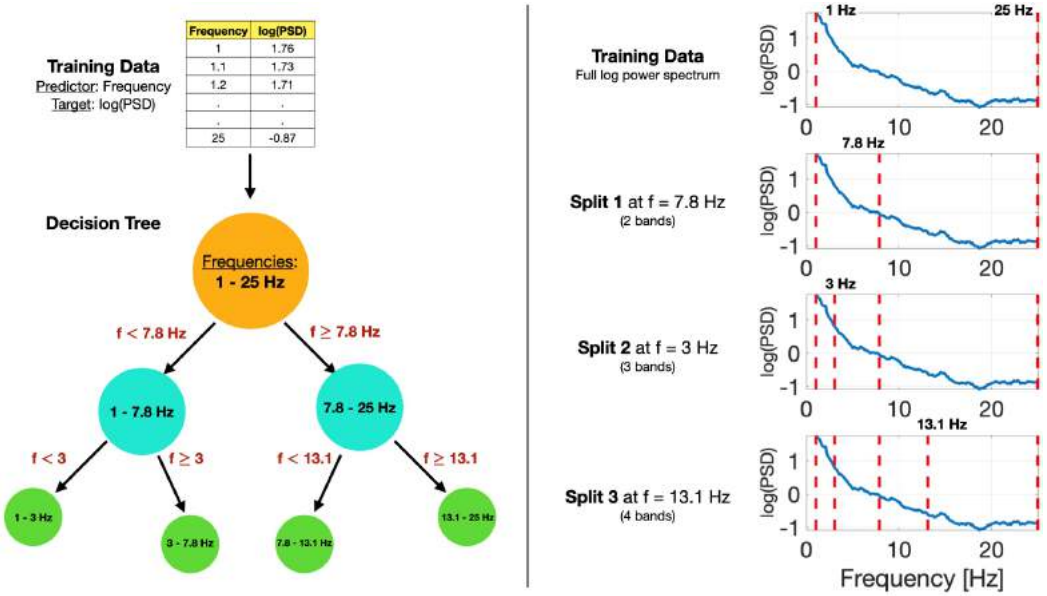


Figure 4. (**Left**) Visual summary of a decision tree partitioning frequency values based on the natural logarithm of the power spectral density. (**Right**) Visualization of decision tree splits of frequency values with power spectra.

We use a single predictor variable (frequency) to estimate a single target variable (natural logarithm of the power spectral density). Any calculation of the power spectral density can be used and plugged into our technique. One such method is described by Welch [4]. Using this technique, for example, the target variable is defined by the following expression.

$$Y_i = \ln \hat{P}(f_i)$$

where, Y_i is the ith element of the target variable array, $\hat{P}$ represents the spectral estimate according to [4], and f_i is the ith element of the predictor variable array i.e., ith frequency value from the EEG signal decomposition.

The decision tree splits frequency values into subgroups and assigns each subgroup a single target value estimation. A greedy search of the decision tree parameter space yields frequency splits that best reproduce target values [25,26]. Thus, through this optimization process, we automatically obtain the optimal member-adjacent frequency bands for a predefined band count.

The band count corresponds to the maximum number of splits used in decision tree training. However, through the use of an AIC-inspired quality score, the proposed method removes the need for manual entry of this quantity. This is discussed further in the next section.

2.4. Quality Score for Band Boundaries

Although decision tree optimization can be leveraged to identify optimal EEG frequency bands, this method requires the number of bands to be predetermined. Instead of choosing a band count manually, here we describe an objective data-driven strategy. The choice of band count is framed as an optimization problem, where we define an objective that can be optimized with respect to the band count.

One choice of objective is the r^2 regression score. In this context, the r^2 value corresponds to how well a set of decision tree-derived EEG band boundaries reproduce an underlying power spectrum. While the decision tree optimization strategy described previously will ensure band boundaries are optimal for a given number of bands, different choices of band count will correspond to different r^2 values. An example of this is shown in Figure 5, where the r^2 regression scores of several different choices of band count are plotted for the same dataset.

However, the r^2 score is a problematic objective choice, since it strictly increases with the number of bands. Therefore, the maximum regression score would correspond to the largest possible number of bands i.e., a frequency "band" for every observed frequency value. One simple solution is to introduce an objective that incorporates both the r^2 regression score and a penalty for the number of bands. This is the goal of popular measures such as the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) [27]. Taking inspiration from AIC, we construct an empirically derived quality score (QS) to help choose a model that balances the best regression score while limiting the number of bands.

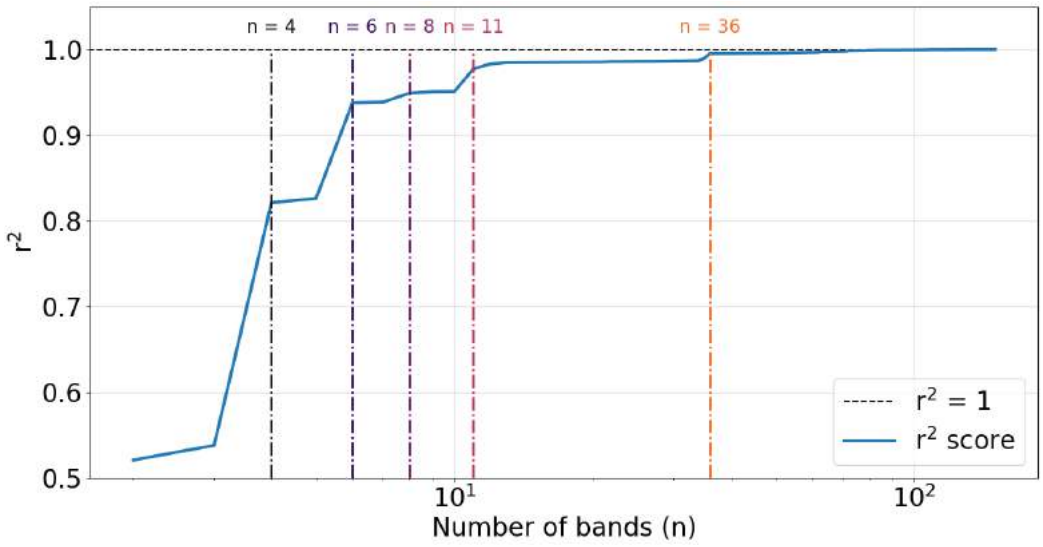


Figure 5. r^2 regression scores plotted against the number of frequency bands included in the decision tree model. The data used to derive these bands and r^2 values is the artificial data described in Case Study 1 in Section 3.1. Colored dashed vertical lines highlight large jumps in r^2 and are labeled by the corresponding number of bands.

AIC is a measure of model quality, where smaller values imply better models [27,28]. It is defined in terms of the maximum value of the likelihood function for the model, L, and the number of parameters in the model, k.

$$AIC = -\log L^2 + 2k$$

The quality score (QS) we employed closely resembles AIC with two modifications. First, in lieu of the squared maximum likelihood value, we used the r^2 regression score. Since r^2 values are between $[0, 1]$, the first term in the QS equation below will be between $[0, \infty)$, however, this range is not very large in practice e.g., for $r^2 \geq 0.135$, the first term is approximately between $[0, 2]$. Second, we divided the second term by N, where N is the maximum number of bands, or equivalently, the total number of observed frequency values. This ensures the second term in the equation below takes values in the range $[0, 2]$.

$$QS = -\log r^2 + 2k/N$$

QS provides a way to compare EEG band boundaries in a way that accounts for both goodness-of-fit and band count. It will typically take values between 0 and 2, where smaller values correspond to better models. By computing the QS for every possible band count, we can choose the best EEG band boundaries as the choice with the smallest QS.

Although QS takes inspiration from AIC, a theoretically grounded quantity, its derivation is empirical, therefore it may not be most suitable for all applications. Furthermore, there are countless other objective choices to optimize band count. The decision tree method described in Section 2.3 is independent of this band count optimization step, and thus can be enhanced by a variety of choices.

2.5. Software Implementation

This two-part technique is implemented using the Sci-Kit learn Python library, a popular and free machine learning software [21]. The decision tree implementation used is the sklearn.tree.DecisionTreeRegressor class. The chosen parameters for the decision tree training are specified in Table 1. A detailed description of each parameter can be found at the Sci-Kit learn documentation: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html (accessed on 6 March 2022).

Table 1. Table specifying the sklearn.tree.DecisionTreeRegressor parameters used in decision tree-based band discovery method. Parameter details can be found at the Sci-Kit learn documentation [21].

Name	Value
criterion	"squared_error"
splitter	"best"
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction	0.0
max_features	None
random_state	None
max_leaf_nodes	Optimized with QS
min_impurity_decrease	0.0
ccp_alpha	0.0

Our code is open-source and publicly available at the GitHub repository: https://github.com/mi3nts/decisiontreeBinning (accessed on 6 March 2022). Although Python is used for our implementation, other statistical software packages can be readily used to implement this method [22–24].

3. Results

In the following subsections, we explore two case studies that apply the proposed data-driven method for EEG frequency band discovery to an artificial and experimental

dataset, respectively. A Python script to reproduce both case studies is freely available at the following GitHub repository: https://github.com/mi3nts/decisiontreeBinning (accessed on 6 March 2022).

3.1. Case Study 1: Artificial Data

As a first demonstration of the method, we produce an artificial EEG power spectrum as shown in Figure 6. The spectrum consists of the characteristic $1/f$ shape for EEG signals with added white noise. A mathematical expression for the artificial power values is given below.

$$P = 1/f + r$$

where, P is the artificially generated power value. f is the frequency value. r is a uniformly distributed random value between 0 and 0.4.

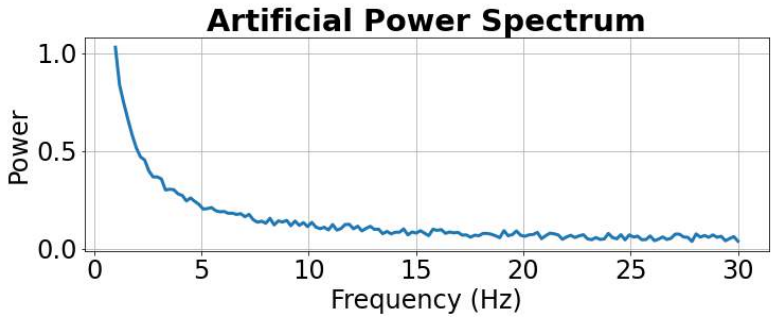


Figure 6. Artificial power spectrum for initial demonstration of data-driven method. The spectrum consists of the characteristic $1/f$ curve for EEG signals with added white noise.

The results of applying decision tree-based band discovery to the artificial power spectrum are shown in Figure 7 for 5 different choices of band count. In the top 5 plots, the true power spectrum is shown as a solid blue line, the decision tree estimated spectrum is plotted as a dashed orange line, and the discovered band boundaries are indicated by dashed vertical red lines. The plots are titled according to the number of bands and r^2 (coefficient of determination) regression score. The r^2 score indicates how well the discovered bands reproduce the true power spectrum. As a comparison, the typical boundaries of the delta, theta, alpha, and beta bands according to the review by [12] are shown at the bottom of Figure 7. Each band is labeled with text. The r^2 score of the standard bands is computed by comparing the average power value within each band with the true values. This value is provided in the plot title.

The greedy search algorithm used in decision tree regression preserves band boundaries when new bands are added. In Figure 7, for example, 7.3 Hz is a band edge in every case (i.e., from 2 bands to 6 bands). It is interesting to note that the discovered 4 bands case is nearly identical to the typical delta, theta, alpha, and beta band boundaries according to [12]. Thus, it may be that the typical band boundaries are a good representation of this characteristic power spectrum.

In Figure 7, as more bands are added, the r^2 regression score increases. A diagrammatic representation of this observation is shown in Figure 5, where model regression scores are plotted against the number of bands. Since there are 150 unique frequency values in this first artificial dataset, the maximum number of bands is 150. Colored dashed vertical lines indicate band choices that exhibit a large jump in the r^2 score.

Since the r^2 score strictly increases with the number of bands, using it as an objective from which to choose the band count would always result in a “band" for every observed frequency value. However, the AIC-inspired quality score (QS) defined in Section 2.4 does not suffer from this issue. This is illustrated in Figure 8, which plots QS against the number of bands. Additionally, the r^2-based fitness term in QS is shown as a dashed blue line, the band count penalty term is plotted as a dashed orange line, and the minimum QS value is

indicated by a yellow star. A minimum QS value is observed at 6 bands, implying the best choice of band count for this spectrum is 6.

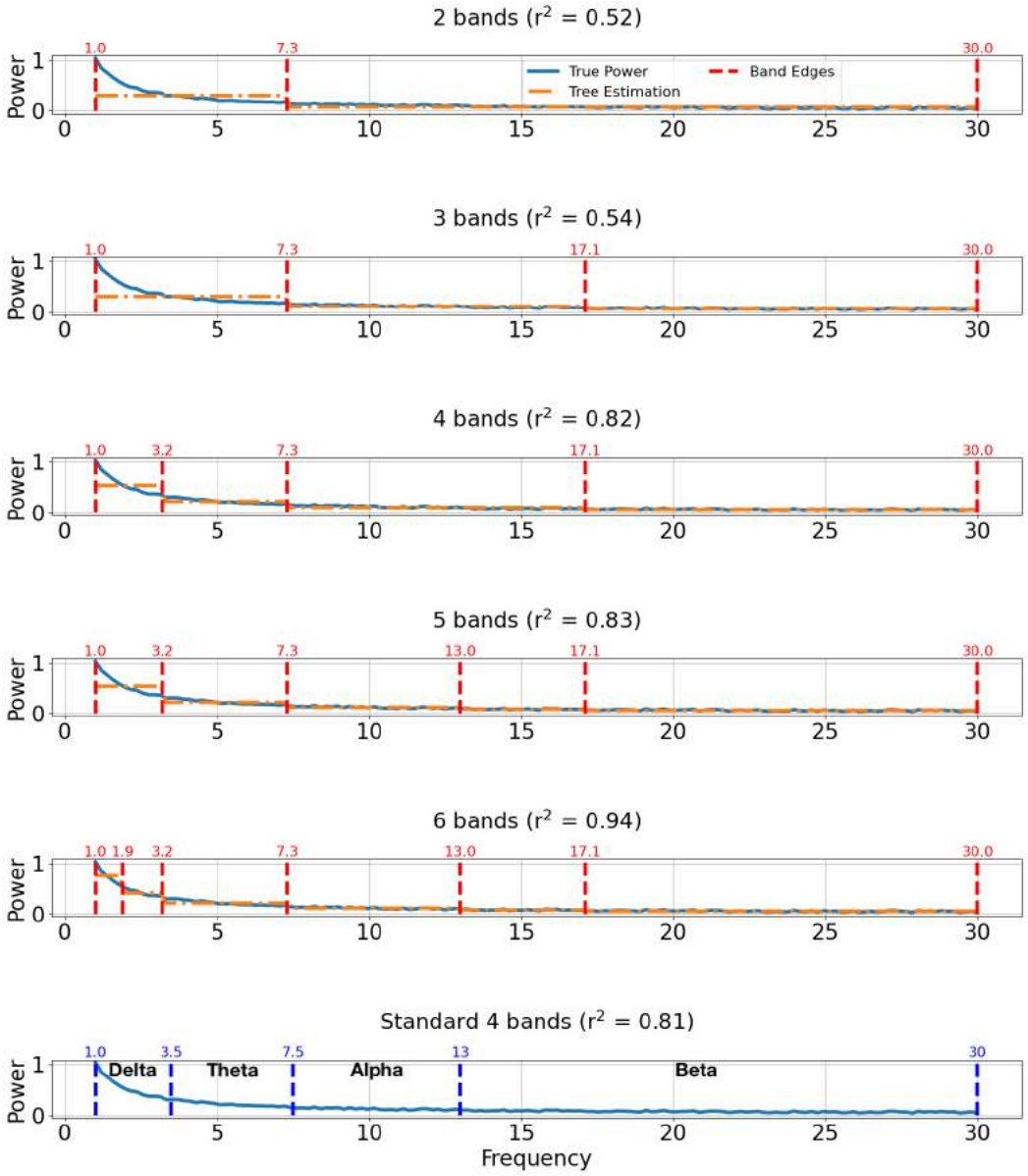


Figure 7. Band comparisons for artificial power spectrum. The true power spectra are plotted with solid blue lines, predicted spectra are plotted with dashed orange lines, and discovered band boundaries are indicated by dashed vertical red lines. The plots are titled according to their number of bands and r^2 regression score. For comparison, typical values of the standard 4 bands (delta, theta, alpha, and beta) according to [12] are shown in the bottom plot along with the true power spectrum plotted again as a solid blue line.

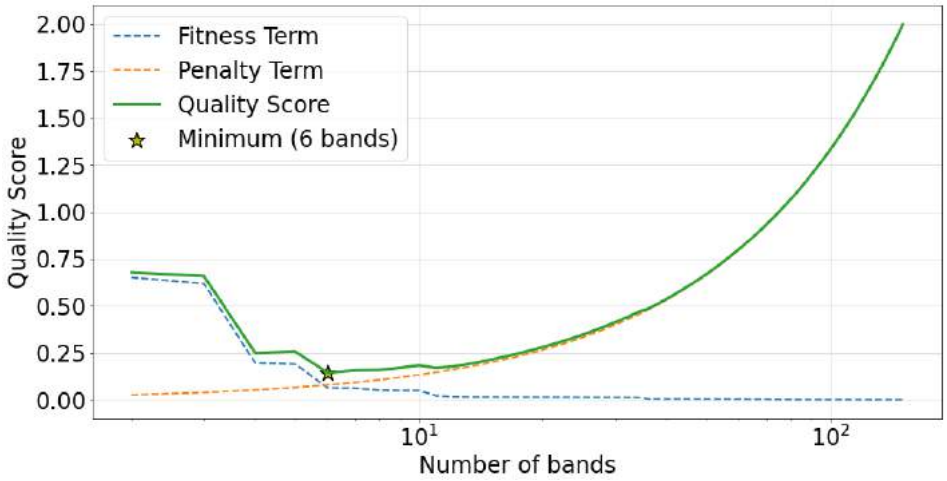


Figure 8. Empirically derived quality score (*QS*) plotted against the number of bands for a case study of an artificially generated power spectrum. The r^2-based fitness term in *QS* is shown as a dashed blue line, the band count penalty term is plotted as a dashed orange line, *QS* is plotted as a green line, and the minimum *QS* value is indicated by a yellow star.

The top plot in Figure 9 outlines the optimal bands based on a quality score (*QS*) minimization strategy. The plot title indicates the number of bands (6), r^2 regression score (0.94), and the quality score of the band definitions (0.14). The bottom plot in Figure 9 similarly outlines the standard bands, titled with the same metrics. The *QS* of the standard bands is computed using the *QS* equation in Section 2.4 with $k = 4$. Although the discovered bands include more parameters, the *QS* is about half of that of the standard bands, thus it is a better characterization of the underlying spectrum based on this objective. Based on the data-driven approach, new band boundaries are discovered that complement the standard delta, theta, alpha, and beta bands by dividing the standard delta and beta bands into two.

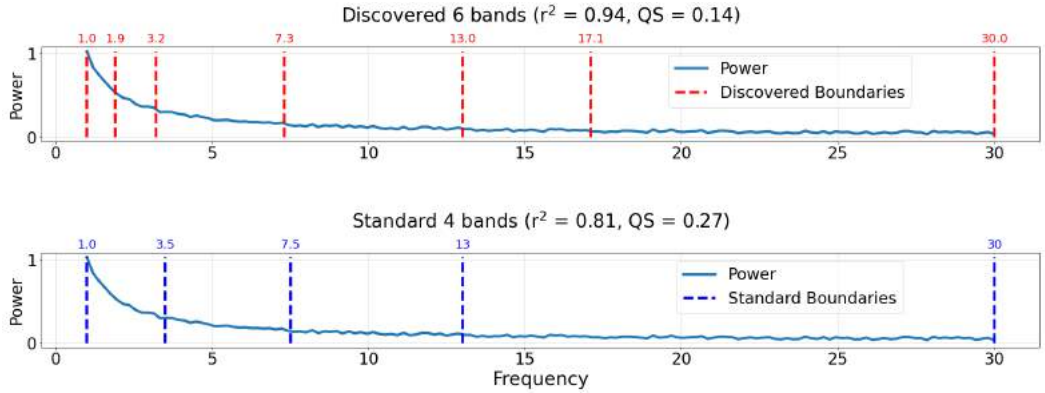


Figure 9. Comparison of discovered and standard bands for the case study of an artificially generated power spectrum. Plots are titled by the number of bands, r^2 regression score, and the quality score of the respective band boundaries. The true power spectrum is plotted as a solid blue line. (**Top**) Discovered bands using the proposed decision tree method employing a minimum quality score (*QS*) technique. Discovered band boundaries are indicated by dashed vertical red lines. (**Bottom**) Typical standard band boundaries are taken from review by Newsom [12]. Standard band boundaries are indicated by dashed vertical dark blue lines.

3.2. Case Study 2: Experimental Data

We evaluate the band discovery method on experimental data from the PhysioNet dataset: EEG During Mental Arithmetic Tasks [29,30]. EEG data were collected monopolarly using the Neurocom EEG 23-channel system (Ukraine, XAI-MEDICA). The electrodes were placed on the scalp according to the International 10/20 montage. Interconnected ear

electrodes were used as the reference. A 30 Hz cut-off frequency high-pass filter and a 50 Hz power line notch filter were used. The data are artifact-free segments of 60 s. In preprocessing, Independent Component Analysis (ICA) was used to eliminate artifacts (eyes, muscles, and cardiac). For this case study, the baseline EEG recording from Subject 00 is used. Occipital electrodes (O1 and O2) are averaged to produce an aggregated occipital EEG signal.

The aggregated occipital time-series signal and its corresponding power spectrum are shown in Figure 10. Welch's method using a Hanning window and a segment length of 1028 estimated the power spectral density of the aggregated signal [4]. Due to the signal preprocessing scheme used here, the power spectrum does not follow the typical $1/f$ shape. Nevertheless, an alpha rhythm peak is observed. Although this experimental power spectrum is characteristically different than the previous artificial spectrum, the EEG bands discovered by our data-driven approach will automatically adapt to it.

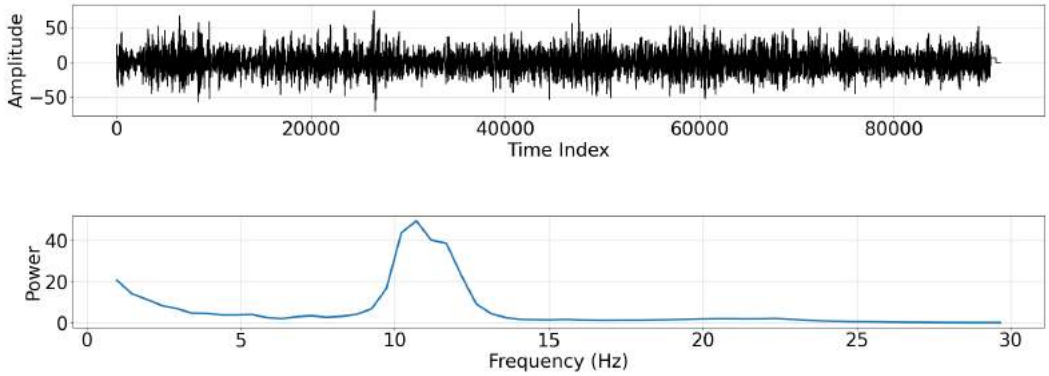


Figure 10. (**Top**) Time series of aggregated occipital EEG signal. (**Bottom**) Power spectral density plotted against frequency for aggregated occipital EEG signal plotted from approximately 1–30 Hz.

We repeat the two-part strategy from Case Study 1. First, we derive band boundaries using the decision tree strategy for every possible choice of band count (i.e., 2 to 60 bands). Second, we use the quality score (QS) to identify the best number of bands. The QS is plotted against the number of bands in Figure 11. The minimum QS value occurs for the 6 bands case.

Figure 12 compares the bands discovered by applying the proposed band discovery strategy to the experimental data (top plot), the optimal bands discovered in Case Study 1 (middle plot), and the standard EEG band boundaries from [12] (bottom plot). The discovered bands from the experimental data (top plot) outperforms the other band choices, with both a significantly higher r^2 score and lower (better) QS. For the present case study, the quality score for the discovered bands was 0.35, compared to scores of 0.86 and 0.8 for the bands discovered in Case Study 1 and standard EEG bands, respectively. Despite the fact that the discovered band boundaries outperformed the standard bands in Case Study 1, when the same boundaries are applied to new experimental data, the standard bands perform better. The poor performance of the bands from Case Study 1 in characterizing this experimental power spectrum highlights the need to tailor EEG bands to specific datasets.

The bands identified from the experimental data isolate spectral features. Specifically, the peak in power spectral density between 10 and 12 Hz is partitioned into a dedicated band. This gives an idea of how the proposed method works. It will tend to learn frequency bands that correspond to peaks in the underlying power spectrum.

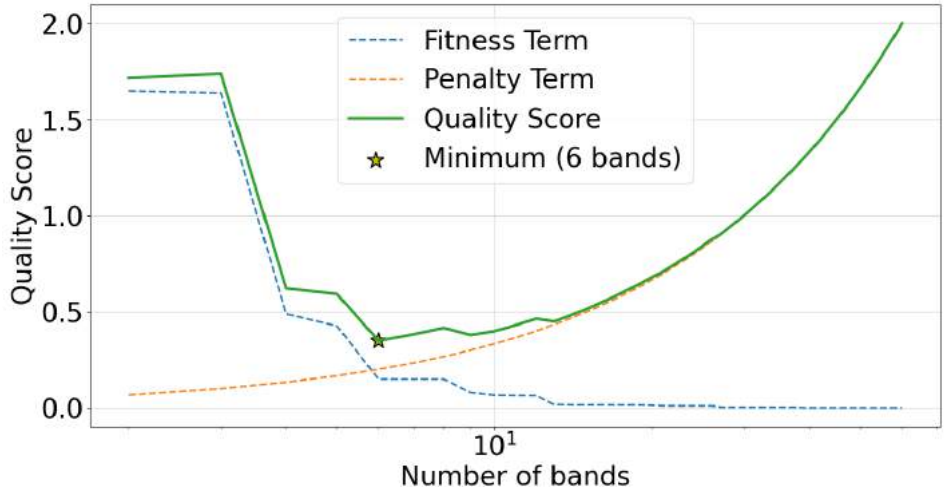


Figure 11. Empirically derived quality score (*QS*) plotted against the number of bands for the case study of experimental EEG data. The r^2-based fitness term in *QS* is shown as a dashed blue line, the band count penalty term is plotted as a dashed orange line, *QS* is plotted as a green line, and the minimum *QS* value is indicated by a yellow star.

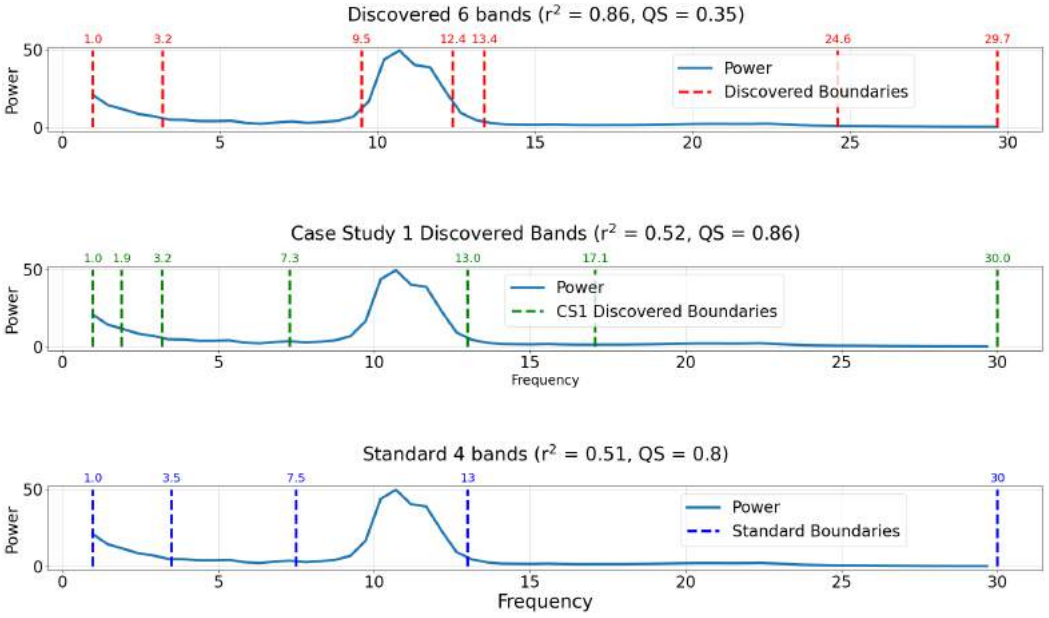


Figure 12. Comparison of discovered and standard bands for the case study of experimental EEG data. Plots are titled by the r^2 regression score and the quality score of the respective band boundaries. The true power spectrum is plotted as a solid blue line. (**Top**) Discovered bands using the proposed decision tree method employing a minimum quality score (*QS*) technique. Discovered band boundaries are indicated by dashed vertical red lines. (**Middle**) Discovered bands derived from artificial power spectrum in Case Study 1. Discovered band boundaries from Case Study 1 are indicated by dashed vertical green lines. (**Bottom**) Typical boundaries of standard bands are taken from a review by Newsom [12]. Standard band boundaries are indicated by dashed vertical dark blue lines.

4. Discussion

This paper outlines a self-supervised method for discovering optimal EEG frequency bands. It differs from previous methods in two important ways. First, band discovery is entirely self-supervised, so an external target variable is not necessary. Second, since the method solely uses a power spectrum, it is capable of handling both single- and multi-channel data across contexts.

The methodology was evaluated by using two case studies. In the first case study, the method was applied to a power spectrum consisting of a $1/f$ shape with white noise added. The discovered bands overlapped with the typical delta, theta, alpha, and beta boundaries. Two additional bands were found within the typical delta and beta frequencies. Despite the larger number of parameters, the discovered bands had a quality score that was nearly half that of the typical ones, thus indicating significantly better performance. In the second case study, the method was applied to a baseline EEG recording from the open-access PhysioNet dataset: EEG During Mental Arithmetic Tasks [29,30]. As in the previous case, the discovered EEG bands significantly outperformed the more conventional boundaries. Additionally, the discovered bands isolated a peak in the power spectral density curve into a dedicated frequency band.

The proposed method has two key strengths. First, the method provides a way to determine frequency bands that are representative of an underlying power spectrum while keeping the number of bands to a minimum. This results in a parameter-free and reproducible approach to the discovery of optimal EEG bands. Second, the method is readily accessible since it is based on decision tree optimization, which has many efficient and ready-to-use implementations [21–24]. Additionally, we made our implementation of the technique open-source and publicly available (https://github.com/mi3nts/decisiontreeBinning (accessed on 6 March 2022)).

Unlike other methods that optimize band boundaries to estimate a particular variable (e.g., disease diagnosis), our approach relies only on power spectral density values for estimation. Although this can be considered a strength, it also has a downside. Since EEG bands are purely based on spectral density curves, their interpretation may not be clear. Consequently, interpretation of bands discovered using this method may require additional effort compared to other approaches.

In Table 2, we present a comparison of different approaches to EEG band discovery. Five previous methods are compared with the one proposed in this article [16–20]. The comparisons are based on four characteristics: supervised, self-supervised, single-channel, and multi-channel. A supervised method is one that uses a target variable to ground the band optimization, such as disease diagnosis, stimulus details, and task type. Conversely, a self-supervised method uses the inherent characteristics of EEG signals to partition bands. Single-channel means the technique can operate on single-channel EEG data, while multi-channel indicates the technique operates on multi-channel EEG data. The uniqueness of our approach resides in the fact that it is a self-supervised approach that works on both single-channel and multi-channel data. Furthermore, it is the first EEG band discovery method that uses the decision tree machine learning framework.

A key feature of the presented approach is that it is agnostic to *how* the input power spectrum is generated, thus it can readily be applied to other types of power spectra (e.g., audio signals, hyperspectral imaging). Hyperspectral imaging, for instance, captures images with layers beyond the standard red, green, and blue. This provides a power spectrum for each pixel of a hyperspectral image. Using the proposed band discovery method, interesting spectral features in hyperspectral images can be detected in a self-supervised way.

Additionally, this method can be applied to other types of input variables. For example, a times series (e.g., heart rate over time) could be used in lieu of a power spectrum. This would result in the discovery of temporal epochs, as opposed to frequency bands.

Table 2. Tabular comparison of different data-driven approaches to EEG band discovery [16–20]. Methods are compared via four characteristics: supervised, self-supervised, single-channel, and multi-channel, shown as columns. Rows indicate the article reference outlining the approaches. The method proposed in this article is shown on the bottom row with the reference name "Proposed Method". An "X" indicates the corresponding method has the listed characteristic.

References	Supervised	Self-Supervised	Single-Channel	Multi-Channel
Elgendi et al. (2011) [17]	**X**		**X**	**X**
Lee et al. (2012) [18]	**X**		**X**	**X**
Magri et al. (2012) [19]	**X**		**X**	**X**
Raza et al. (2015) [20]	**X**			**X**
Cohen (2021) [16]		**X**		**X**
Proposed Method		**X**	**X**	**X**

5. Conclusions

EEG serves as a window to underlying neural processes. Spectral analysis of EEG examines the oscillations in electric potentials arising from the brain. Despite the widespread use of established delta, theta, alpha, and beta bands for EEG, their boundaries vary widely across studies, which may be a result of variations in experimental details and participant differences. This motivates the use of objective and data-driven approaches to EEG band discovery.

In this work, we leveraged the readily available optimization of a decision tree for regression to discover EEG bands most appropriate for a given dataset and a predetermined number of bands. The best choice of band count was then determined using an AIC-inspired quality score. We applied the presented method to both artificial and open-access experimental data. Discovered bands isolated spectral features into dedicated bands and outperformed the standard band definitions. Data-driven EEG band discovery may provide new indices of neural activity which can adapt to a variety of experimental and subject characteristics.

Author Contributions: Conceptualization, S.T. and J.W.; methodology, S.T., J.W, and D.J.L.; software, S.T.; validation, S.T. and J.W.; formal analysis, S.T.; investigation, S.T.; data curation, S.T.; writing—original draft preparation, S.T.; writing—review and editing, S.T., J.W., B.A.F., A.S., D.J.L.; visualization, S.T.; supervision, D.J.L.; project administration, S.T. and D.J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the following grants: 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, award number W81XWH-18-1-0400). 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. Support from the University of Texas at Dallas Office of Sponsored Programs, Dean of Natural Sciences and Mathematics, and Chair of the Physics Department are gratefully acknowledged. The authors acknowledge the OIT-Cyberinfrastructure Research Computing group at the University of Texas at Dallas and the TRECIS CC* Cyberteam (NSF 2019135) for providing HPC resources that contributed to this research (https://utdallas.edu/oit/departments/circ/, accessed 22 February 2022).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Artificial dataset used in this work can be generated using software provided at the GitHub repository: https://github.com/mi3nts/decisiontreeBinning (accessed on 6 March 2022). Experimental data is from the PhysioNet dataset: EEG During Mental Arithmetic Tasks https://physionet.org/content/eegmat/1.0.0/ (accessed on 6 March 2022) [29,30].

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

EEG	Electroencephalography
BIC	Bayesian Information Criterion
AIC	Akaike Information Criterion

References

1. Mulert, C.; Lemieux, L. *EEG-fMRI: Physiological Basis, Technique, and Applications*; Springer Science & Business Media: Cham, Switzerland, 2010; pp. 1–539. [CrossRef]
2. Jackson, A.F.; Bolger, D.J. The neurophysiological bases of EEG and EEG measurement: A review for the rest of us. *Psychophysiology* **2014**, *51*, 1061–1071. [CrossRef] [PubMed]
3. Cohen, M.X. Where Does EEG Come From and What Does It Mean? *Trends Neurosci.* **2017**, *40*, 208–218. [CrossRef] [PubMed]
4. Welch, P.D. 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. [CrossRef]
5. Louis, E.K.S.; Frey, L.C. *Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants*; American Epilepsy Society: Chicago, IL, USA, 2016; pp. 1–95.
6. Silva, F.L.D. EEG: Origin and Measurement. In *EEG-fMRI: Physiological Basis, Technique, and Applications*; Springer Science & Business Media: Cham, Switzerland, 2009; pp. 19–38. [CrossRef]
7. Mills, C.; Fridman, I.; Soussou, W.; Waghray, D.; Olney, A.M.; D'Mello, S.K. Put your thinking cap on: Detecting cognitive load using EEG during learning. In *LAK '17: Proceedings of the Seventh International Learning Analytics & Knowledge Conference*; Association for Computing Machinery: New York, NY, USA, 2017; pp. 80–89. [CrossRef]
8. Friedman, N.; Fekete, T.; Gal, K.; Shriki, O. EEG-based prediction of cognitive load in intelligence tests. *Front. Hum. Neurosci.* **2019**, *13*, 191. [CrossRef] [PubMed]
9. Kumar, N.; Kumar, J. Measurement of Cognitive Load in HCI Systems Using EEG Power Spectrum: An Experimental Study. *Procedia Comput. Sci.* **2016**, *84*, 70–78. [CrossRef]
10. De Medeiros Kanda, P.A.; Anghinah, R.; Smidth, M.T.; Silva, J.M. The clinical use of quantitative EEG in cognitive disorders. *Dement. Neuropsychol.* **2009**, *3*, 195. [CrossRef] [PubMed]
11. Cassani, R.; Estarellas, M.; San-Martin, R.; Fraga, F.J.; Falk, T.H. Systematic review on resting-state EEG for Alzheimer's disease diagnosis and progression assessment. *Dis. Mark.* **2018**, *2018*, 5174815. [CrossRef] [PubMed]
12. Newson, J.J.; Thiagarajan, T.C. EEG Frequency Bands in Psychiatric Disorders: A Review of Resting State Studies. *Front. Hum. Neurosci.* **2019**, *12*, 521. [CrossRef] [PubMed]
13. Li, M.; Lu, B.L. Emotion classification based on gamma-band EEG. In Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009, Minneapolis, MN, USA, 3–6 September 2009; pp. 1323–1326. [CrossRef]
14. Aljribi, K.F. A Comparative Analysis of Frequency Bands in EEG Based Emotion Recognition System. In *ACM International Conference Proceeding Series*; Association for Computing Machinery: New York, NY, USA, 2021. [CrossRef]
15. Gannouni, S.; Aledaily, A.; Belwafi, K.; Aboalsamh, H. Emotion detection using electroencephalography signals and a zero-time windowing-based epoch estimation and relevant electrode identification. *Sci. Rep.* **2021**, *11*, 7071. [CrossRef]
16. Cohen, M.X. A data-driven method to identify frequency boundaries in multichannel electrophysiology data. *J. Neurosci. Methods* **2021**, *347*, 108949. [CrossRef]
17. Elgendi, M.; Vialatte, F.; Cichocki, A.; Latchoumane, C.; Jeong, J.; Dauwels, J. Optimization of EEG frequency bands for improved diagnosis of Alzheimer disease. In Proceedings of the 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Boston, MA, USA, 30 August–3 September 2011; Volume 2011, pp. 6087–6091. [CrossRef]
18. Lee, C.; Jung, J.; Kwon, G.; Kim, L. Individual optimization of EEG channel and frequency ranges by means of genetic algorithm. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, San Diego, CA, USA, 28 August–1 September 2012; pp. 5290–5293. [CrossRef]
19. Magri, C.; Mazzoni, A.; Logothetis, N.K.; Panzeri, S. Optimal band separation of extracellular field potentials. *J. Neurosci. Methods* **2012**, *210*, 66–78. [CrossRef]
20. Raza, H.; Cecotti, H.; Prasad, G. Optimising frequency band selection with forward-addition and backward-elimination algorithms in EEG-based brain-computer interfaces. In Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 12–17 July 2015. [CrossRef]

21. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
22. MathWorks. Fitrtree. 2021. Available online: https://www.mathworks.com/help/stats/fitrtree.html (accessed on 6 March 2022).
23. DataCamp. rpart: Recursive Partitioning and Regression Trees. 2021. Available online: https://www.rdocumentation.org/packages/rpart/versions/4.1.16/topics/rpart (accessed on 6 March 2022).
24. DecisionTree.jl Documentation. 2021. Available online: https://docs.juliahub.com/DecisionTree/pEDeB/0.10.8/autodocs/ (accessed on 6 March 2022).
25. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Routledge: New York, NY, USA, 2017; pp. 1–358. [CrossRef]
26. Kotsiantis, S.B. Decision trees: A recent overview. *Artif. Intell. Rev.* **2011**, *39*, 261–283. [CrossRef]
27. Salkind, N. Bayesian Information Criterion. In *Encyclopedia of Measurement and Statistics*; SAGE Publications: Thousand Oaks, CA, USA, 2013; pp. 1–3. [CrossRef]
28. Akaike, H. A New Look at the Statistical Model Identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
29. Zyma, I.; Tukaev, S.; Seleznov, I.; Kiyono, K.; Popov, A.; Chernykh, M.; Shpenkov, O. Electroencephalograms during Mental Arithmetic Task Performance. *Data* **2019**, *4*, 14. [CrossRef]
30. Goldberger, A.L.; Amaral, L.A.N.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation* **2000**, *101*, e215–e220. [CrossRef] [PubMed]

MDPI

Article

Evaluating Physiological Performance Predictors in High-Stress, Live-Fire Scenarios: A Pilot Study

Firstname Lastname [1,†,‡], Firstname Lastname [2,‡] and Firstname Lastname [2,*]

[1] Affiliation 1; e-mail@e-mail.com
[2] Affiliation 2; e-mail@e-mail.com
* Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials) +xx-xxxx-xxx-xxxx (F.L.)
† Current address: Affiliation 3.
‡ These authors contributed equally to this work.

Abstract: Autonomic responses, such as the stress response, are automatic physiological mechanisms meant to keep us safe. They provide a window into underlying cognitive and physiological processes. The processing of novel life-threatening stimuli consumes cognitive resources which can impede decision making and performance. A deeper and more detailed understanding of autonomic physiological responses in the context of high-stress live-fire scenarios offers the possibility of performance improving interventions and preserving human life. In this pilot study we captured a comprehensive picture of the participant's physical and cognitive status in the context of hyper-realistic live-fire training scenarios and used it to predict performance. The high-dimensional space of biometric markers lends itself to the use of machine learning to develop objectively optimized empirical models of performance.

Keywords: Machine Learning; Holistic Sensing; Human Performance; Electroencephalography (EEG)

1. Introduction

Warfighters, federal agents, law enforcement personnel, and first responders routinely go into harm's way to protect and serve. Deadly threat encounters frequently lead to physiological effects such as elevated heart rate and respiration rate, pupil dilation, auditory exclusion, loss of dexterity, trembling, and an adrenaline increase. In more extreme circumstances, freezing can also occur, as well as the loss of bladder control and vocal control. These responses typically occur very rapidly over a few seconds. For this reason, training is essential to develop appropriate skills, tools, and behaviors when faced with such extreme physiological and psychological stressors. It is essential that a framework is provided to encounter these natural, yet unconscious, extreme responses in a safe environment, and then to provide detailed insights to help learn how to manage them. Training can enable both the ability to appropriately modulate responses in the face of lethal threats while also preserving innocent lives and minimizing unintended collateral damage.

The purpose of this pilot study is to evaluate physiological interactions with human performance in hyper-realistic, live-fire training scenarios. Participants performed complex scenarios based on real-life events while equipped with a suite of physiological sensors. Performance was assessed by two expert evaluators on a scale of 0 to 4 for the following 8 dimensions: Decision making, Timeliness, Cognitive shifting, Use of empathy, Tactical performance, Body language, Self-assessment, and Rationality.

The relationship between physiological quantities and performance are assessed via objectively optimized, empirical machine learning models which predict expert evaluator performance scores from physiological data alone. Models are evaluated based on their accuracy and provide rankings of the most important physiological predictors in estimating performance scores. High fidelity empirical models may provide clues to key physiological

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Journal Not Specified* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

indicators of performance. Such insights enable the development of objective performance scoring systems and can further inform interventions to enhance rapid decision making in high-stress, life-threatening situations.

2. Materials and Methods

This prototype study brings together three key elements for the provision of empirical machine learning models of human performance in high-stress, live-fire scenarios based solely on physiological recordings. These elements include: the Troysgate Training System, Expert Evaluation of Performance, and Holistic Biometric Sensing Suite. These are described in the following subsections followed by an overview of the model development procedure.

2.1. Troysgate Training System

High-stress, live-fire scenarios were simulated with the Troysgate training system. Troysgate creates a highly effective and realistic training environment that places the participant in a "face-to-face," close quarters, deadly conflict situation with a combatant or adversary (US Patent US9453711B2). The participant and role player are safely positioned adjacent to one another and separated by a ballistic containment barrier which provides protection from accidental weapon discharge. A large mirrored reflective screen is positioned several feet in front of participants, so that the participant and role player can easily track each other's actions.

2.2. Expert Evaluation of Performance

Performance was assessed by two independent expert evaluators. Both evaluators are retired army special forces and followed a shared grading scale guidelines. One evaluator was present in the room during each scenario, the second evaluator watched scenarios in real-time via a security camera system. Performance is evaluated at two predetermined points within each scenario. These points are referred to as impact points and consist of critical events in the interaction for which the participant must make a decision (e.g. shoot or don't shoot).

Performance scores were graded on an integer scale of 0 to 4 for the following eight performance dimensions: Decision making, Timeliness, Cognitive shifting, Use of empathy, Tactical performance, Body language, Self-assessment, and Rationality. These performance dimension were developed specifically for this pilot study. More details on how these measures are assessed is provided in Appendix A.

2.3. Holistic Biometric Sensing Suite

During scenario simulations, participants were equipped with a holistic biometric sensing suite. This sensor package aims to comprehensively and continuously capture the physiological and cognitive responses of the participant, without limiting actions, movements, or decision making. The goal is to gather the maximum amount of information with the least disruption of normal behaviors.

The biometric sensors measure the autonomic physiological responses of the participants by passively collecting data on 115 biometric parameters (at 500 Hz: 64 electrode EEG, ECG, GSR, blood oxygen, skin temperature, 6-axis IMU (gyro and accelerometer), 100 Hz: Eye-tracking glasses). After processing this information, we are left with about 16,500 variables for analysis. Additionally, the eye tracking glasses record a point-of-view video and capture video of each pupil. These videos are also processed for additional variables. Sensor recording units and other devices are organized in a backpack worn by the participant that all together weighs less than 10 lbs.

This holistic biometric sensing suite integrates two independent sensing systems (Figure 1). Eye tracking is recorded 100 times a second using the Tobii Pro Glasses 2. All other biometric data are measured 500 times a second using the Cognionics Mobile-64 and AIM2 systems. The most important detail in bringing these two systems together is ensuring

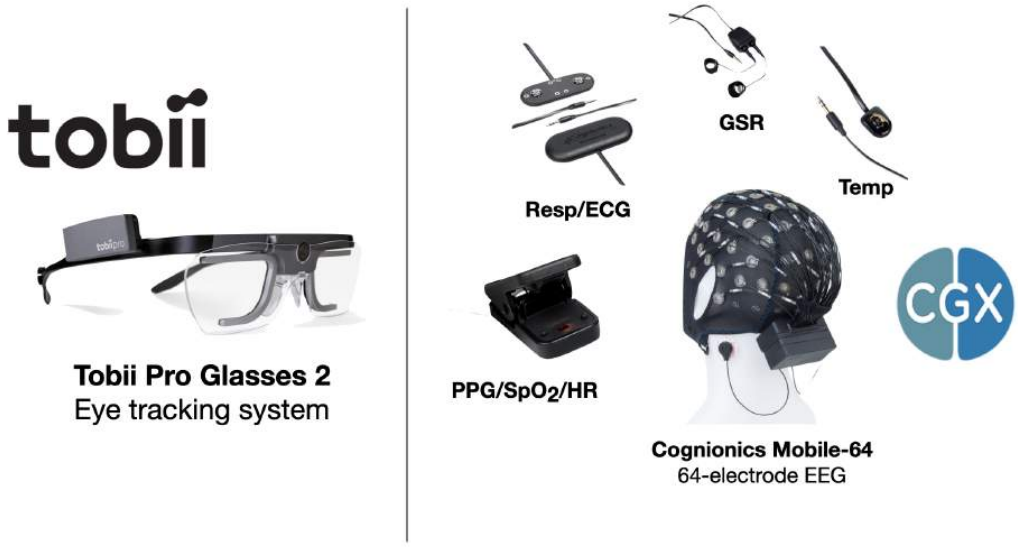


Figure 1. Biometric sensing systems. (**Left**) Tobii Pro Glasses 2 eye tracking system. This instrument performs eye tracking data, pupillometry, and provides two videos streams of the participant's POV and eyes, respectively. (**Right**) Cognionics Mobile-64 and AIM2 systems. Sensing suite includes 64-electrode EEG, PPG which measures SpO_2 and HR, respiration/ECG sensors, GSR, and temperature probe.

data are stored to a common time index. Both hardware and software technologies are leveraged to achieve this goal. The Cognionics wireless trigger receives timing signals from the Tobii Pro Glasses 2, which are used to align data from the two systems. Furthermore, eye tracking and pupillometric variables are up-sampled to match the 500 Hz sampling rate via a linear interpolation after taking care of error values.

2.4. Supervised Machine Learning

Supervised machine learning makes it possible for computers to learn by example. Thus, more examples lead to better models. The present pilot analysis contained only 28 instances from which the supervised machine learning framework could learn. Although the data is limited, we map an analysis protocol that can be leveraged for future studies and provide preliminary results.

During model training, predictor variables are mapped to target variables. This results in an empirical machine learning model that can estimate target variables given new input data. A subset of records is excluded from the training process to serve as a validation dataset that tests the model's performance with new data. In this case, 20% of all available data records are held back for validation. With the limited dataset size, only 6 records are included in the validation dataset. The small size of the validation set may not represent the underlying data distribution, and so may not yield statistically significant results.

The supervised machine learning approach used 329 predictor variables to estimate 27 different target variables, one model for every target, from 28 data records. Each model has a single target variable. These target variables were the experienced instructors assessments of a impact point for a specific performance dimension.

Twenty-seven models result from one model for each of the eight grades plus their sum, as assigned by evaluator 1, evaluator 2, and their average. Of the 27 models, 11 are developed using an ensemble of decision trees for regression and the remaining 16 are derived using an ensemble of decision trees for classification. The key difference between these types of models lies in their target variables. Namely, regression models have targets with continuous values. That is, values that span a number line. Conversely, classification

models have discrete targets. Meaning they aim to predict values that are distinct and do not sit on a common continuous axis e.g. cat, dog, pig, etc.

A target which is the average of the two evaluators' performance grades or the total of the eight performance grades is treated as a continuous target for regression. All other grades are evaluated as discrete targets for classification. Even though evaluator scores are presented on a scale, they are treated as discrete labels in the development of the 16 classification models. This means that a score of 0 and 1 differs just as much as a score of 0 and 4.

2.4.1. Data Preparation

During key scenario impact points, participant performance is assessed by expert evaluators. For each impact point, biometric variables are averaged over a time frame of 10 seconds. Further, these biometrics averages are scaled to the individual participant's baseline averages to account for individual differences. This scaling is calculated as follows.

$$\bar{x}_{scaled} = \frac{\bar{x} - \bar{x}_{baseline}}{\bar{x}_{baseline}}$$

Where, $\bar{x}_{scaled}$ represents the average value of a biometric variable over a 10 second epoch spanning a impact point scaled to the average baseline of the same biometric variable. $\bar{x}$ represents the unscaled average value of a biometric variable over a 10 second epoch spanning a impact point. And, $\bar{x}_{baseline}$ is the average value of a biometric variable over the course of a baseline recording.

3. Results

The objective of this analysis was to determine the key biometric effects on cognitive performance during high-stress live-fire scenarios. This analysis consists of two main steps. In the first step, 27 different models of performance are trained using supervised machine learning. Next, the fidelities of these models are evaluated and predictor importance rankings are obtained.

3.1. Model Evaluation

Each of the 27 models are evaluated via two key data products. First is model fidelity. For regression models this is quantified by the correlation coefficient between the true evaluator scores and the model predicted scores. For classification models this is assessed by the percentage of correct evaluator score classifications. The second data product is a predictor importance ranking. In these models, predictor importance is based on the information gained about a target performance variable when conditioning on a given predictor variable.

The participant performance dimensions that could best be reproduced using the predictor variables and the supervised machine learning framework were the mean scores for Cognitive Shifting, Rationality, and Self-assessment. The left plot in Figure 2 shows the ranking of model accuracy for the regression models. This indicates the biometric predictor variables are most relevant in identifying performance along these dimensions. These dimensions were graded based on the following criteria.

Cognitive Shifting – Cognitive shifting is the ability to shift from one tactic of engagement to another. Tactic in this sense would be referring to the use of an empathetic, compassionate tone to strong verbal commands. The ability to shift as the scenario evolves or as use of a particular tactic does not achieve desired effect is what is being assessed in this metric.

Rationality – Rationality is scored after the scenario during the debrief when the participant discusses what occurred. This criterion refers to the validity of the participant's rationale for their decision making during the scenario in terms of supporting their assessment of how they were able to intervene. For example, not just the decision to shoot a role player that presented a threat but why and when they made that decision.

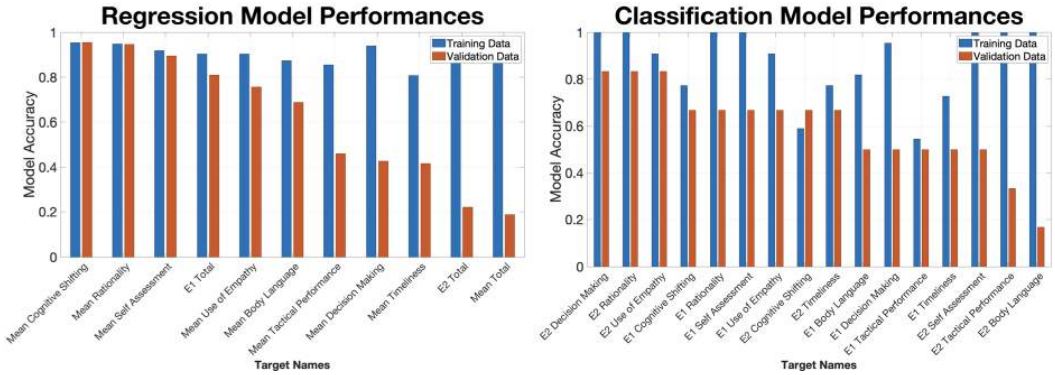


Figure 2. Assessment of 27 performances models. (**Left**) Accuracy of 11 performance models based on ensemble of decision trees for regression. Bar heights indicate squared correlation coefficients for each model. Training dataset performance is shown in blue and validation data performance in orange. (**Right**) Accuracy of 16 performance models based on ensemble of decision trees for classification. Bar heights indicate ratio of number of correct classifications to the total number of records. Training dataset performance is shown in blue and validation data performance in orange.

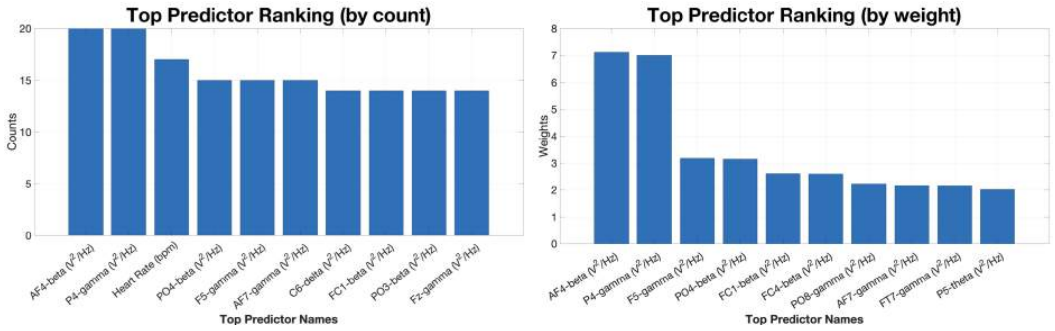


Figure 3. Ranking of top biometric predictor variables in reproducing expert evaluator scores. (**Left**) Top biometric predictors based on the number of performance models in which the respective biometric variable had a non-zero importance. (**Right**) Top biometric predictors based on the aggregated importance weight of each biometric variable across all 27 performance models.

Self-assessment – Self-assessment is scored after the scenario during the debrief when the participant discusses what occurred and their assessment of how they feel about their performance. Additionally, the Self-assessment also includes the level of detail the participant is able to articulate regarding what occurred as an indication of the connection between their executive function and recall of sensory input.

3.2. Top Biometric Predictors

The top three biometric predictors of performance were the baseline scaled beta band (13 – 25 Hz) power spectral density of the AF4 electrode recording (AF4-beta), the baseline scaled gamma band (25 – 75 Hz) power spectral density of the P4 electrode recording (P4-gamma), and heart rate values scaled to baseline. The biometric variables AF4-beta and P4-gamma appear in the top predictor importance rankings based on both count and weight (Figure 3). The ranking based on count indicates the number of performance models in which each biometric predictor had a non-zero importance, while the weight-based ranking shows the sum of each biometric variable's importance across all 27 models. The scaled heart rate value appeared in the top 3 of the count-based ranking but not the weight-based one. This may suggest that heart is a universally relevant predictor variable, however other biometrics play a larger role when it comes to individual models.

EEG signals arise from neurons with synchronized activity. Additionally, local neural orientations and relative location to head surface impact captured signals [1]. These signals can be decomposed into frequency bands e.g. beta (13 – 25 Hz) and gamma (25 – 75 Hz).

That is, frequency bins that reflect characteristic signal oscillations. Generally speaking, high frequency components correspond to higher neural firing rates and thus greater neural activity.

The total beta (13 – 25 Hz) band power for the AF4 electrode. The AF4 electrode is located near the front right part of the participant's head over the frontal lobe. It records activity from Brodmann area 9 (ba09). This large area of the frontal cortex is associated with overriding automatic reactions [2], evaluating intention [3], theory of mind [4], working memory [5–7], auditory attention [8], planning [9], and recognizing the emotions of others [10].

The total gamma (25 – 75 Hz) band power for the P4 electrode. The P4 is near the back right part of the head over the parietal lobe. This electrode corresponds to activity arising from Brodmann area 39 (ba39). Although there is less research on the function of ba39 compared to ba09, it has been associated with functions such as overriding automatic reactions [2], attention and spatial cognition, reasoning, and social cognition [11].

4. Discussion

Authors should discuss the results and how they can be interpreted from the perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

5. Limitations & Future Works

The power of machine learning is its ability to parse high-dimensional and high volume datasets. A comprehensive dataset is a key ingredient to successful machine learning models and analyses. Although we procured a comprehensive snapshot of a participant's biometric status and performance evaluation, the data volume, i.e. the number of data records, was limited in this pilot project. The lack of EEG data collection resulting from participant hair length and other equipment malfunctions led to fewer comprehensive recordings, down from 48 to 28 impact points.

A high volume dataset helps ensure the information from which the machine learning models learn is a diverse and representative sample of the population of interest. Therefore, it should be noted that the findings of this study should not be taken as fact, but rather as hints for future research. A deeper investigation with a larger sample size will enable us to establish statistical significance and develop high fidelity models of performance. This would facilitate the evaluation of key predictors for specific performance dimensions and thus provide an opportunity for targeted interventions e.g. using key predictors of decision making to inform training and tactical protocols.

6. Conclusions

This study outlines an experimental paradigm for assessing the interactions between participants' performance in high-stress, live-fire training scenarios and a holistic suite of biometric sensors. Despite limited data in this pilot study, preliminary results were obtained using both unsupervised and supervised machine learning frameworks. With unsupervised machine learning, two composite indices were derived that vary with a participant's cognitive load and focus, respectively. A supervised machine learning approach was then used to estimate the performance scores of expert evaluators using only biometric data. Performance dimensions best captured by the pilot dataset were: Cognitive shift, rationality, and self-assessment. Key biometric predictors in estimating performance scores were the total beta (13 – 25 Hz) band power for the AF4 electrode measuring activity in Brodmann area 9, the total gamma (25 – 75 Hz) band power for the P4 electrode measuring activity in Brodmann area 39, and heart rate.

7. Patents

This section is not mandatory, but may be added if there are patents resulting from the work reported in this manuscript.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: Please add: "This research received no external funding" or "This research was funded by NAME OF FUNDER grant number XXX." and and "The APC was funded by XXX". Check carefully that the details given are accurate and use the standard spelling of funding agency names at https://search.crossref.org/funding, any errors may affect your future funding.

Institutional Review Board Statement: In this section, you should add the Institutional Review Board Statement and approval number, if relevant to your study. You might choose to exclude this statement if the study did not require ethical approval. Please note that the Editorial Office might ask you for further information. Please add "The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving humans. OR "The animal study protocol was approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving animals. OR "Ethical review and approval were waived for this study due to REASON (please provide a detailed justification)." OR "Not applicable" for studies not involving humans or animals.

Informed Consent Statement: Any research article describing a study involving humans should contain this statement. Please add "Informed consent was obtained from all subjects involved in the study." OR "Patient consent was waived due to REASON (please provide a detailed justification)." OR "Not applicable" for studies not involving humans. You might also choose to exclude this statement if the study did not involve humans.

Written informed consent for publication must be obtained from participating patients who can be identified (including by the patients themselves). Please state "Written informed consent has been obtained from the patient(s) to publish this paper" if applicable.

Data Availability Statement: In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Please refer to suggested Data Availability Statements in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics. If the study did not report any data, you might add "Not applicable" here.

Acknowledgments: In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest: Declare conflicts of interest or state "The authors declare no conflict of interest." Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript, or in the decision to publish the results must be declared in this section. If there is no role, please state "The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results".

Sample Availability: Samples of the compounds ... are available from the authors.

Abbreviations

The following abbreviations are used in this manuscript:

Appendix A Grading Scale Guidelines

1. Decision making – across this scale the determination is clear decisions on how to address the scenario such as how to orient to the environment appropriately and/or which role players to address, what tone to use, at what point a threat is great enough to warrant presenting the pistol. This is an overall assessment of what is evidenced of their decision making. A 0 on this scale would be no evidence of clear decision making that results in any form of action, essentially just waiting for the role players to reach a point at which there is an obvious need for the use of lethal force. A "4" on this scale would be clear decisions made at impact points that demonstrate use of executive function appropriate to the evolving situation.

- 0 – no demonstrable decision making, aloof
- 1 – little decision making or only cursory attempts to address the evolving situation
- 2 – some decision making but not appropriately addressing the evolving situation
- 3 – good decision making with continued adjustments to the situation
- 4 – clear, strong decision making appropriately oriented throughout the entire scenario

2. Timeliness – the participant is challenged with a rapidly changing scenario and, as such, must address all nuances in a rapid fashion or they will not have the opportunity to impact the situation. This has some potential overlap with Tactical Performance but only in the aspect of the timing of decisions related to tactical orientation. For example, at what point to draw the pistol and have in a ready position or at what point to engage a target. This can also address when to interject verbal commands or conversation.

- 0 – very slow to orient to the situation or no real measurable responses to occurrences
- 1 – slow to orient to the situation and actions are reactive rather than proactive
- 2 – moderate speed in actions that begin to provide opportunities to impact scenario
- 3 – rapid responses that demonstrate potential to impact the scenario
- 4 – very responses that demonstrate proactive potential

3. Cognitive Shifting – Cognitive shifting is the ability to shift from one tactic of engagement to another. Tactic in this sense would be referring to the use of an empathetic, compassionate tone to strong verbal commands. The ability to shift as the scenario evolves or as use of a particular tactic does not achieve desired effect is what is being assessed in this metric.

- 0 – No change in approach throughout scenario, i.e. only strong verbal commands
- 1 – Little change in approach with minor attempts at shifting
- 2 – Some change in tactics but not appropriately enough to impact behavior
- 3 – Changes in tactics that have some potential for impacting behavior
- 4 – Dynamic change in tactics with clear potential for impacting behavior

4. Use of Empathy – Use of empathy refers to the participant's demonstrated ability to engage the role players from an orientation that relates to their context of experience within the scenario. For example, an angry, scared role player who is expressing outrage would be successfully engaged by the participant expressing validation for their anger and fear.

- 0 – No demonstrated use of empathy
- 1 – Minor use of empathy but attempts are not authentic or believable
- 2 – Use of empathy but attempts are not truly authentic
- 3 – Use of empathy is appropriate and authentic
- 4 – Use of empathy is effective, authentic and potentially impactful throughout scenario

5. Tactical Performance – this refers to the participants' tactical approach to the scenario in terms of position, proper use of the pistol, ability to recognize and address threats, and

- 0 – Participant makes gross tactical errors such as unsafe operation of a pistol or shooting a role player that doesn't present a threat
- 1 – Participant makes multiple tactical errors and/or improperly engages targets (wrong target, before target is a true threat etc)
- 2 – Participant makes an obvious tactical error and/or doesn't orient to the scenario appropriately (too aggressive or too aloof/lackadaisical)
- 3 – Participant appropriately tactically orients to the scenario and makes no obvious errors
- 4 – Participant demonstrates a strong, appropriate tactical orientation to the scenario and decisively, properly engages threats presented

6. Body Language – The posture, orientation, hand gestures, and movement throughout the scenario conveys a great deal of communication visible to the role players. It is important to recognize what is being communicated and if it has potential for positive or negative impact. Is the participant fidgety, nervous, or in a defensive posture? Is the participant in an open posture that conveys calm, confidence or is the participant bladed off, shoulders hunched up in an aggressive stance? How are this positions and movements relating to the scenario? Are they appropriate, inappropriate or conveying communication that will have beneficial or disruptive impact?

- 0 – Participant is continuously excessively nervous, aggressive or aloof relative to the situation (continuous fidgeting or crossed legs with hands on the "stop" wall)
- 1 – Participant is continuously excessively nervous, aggressive or aloof relative to the situation
- 2 – Participant gestures excessively or stiffly in attempting to communicate such that he/she appears oddly affected
- 3 – Participant's body language is appropriate for the scenario and conveys an appropriate orientation to the scenario
- 4 – Participant is clearly aware of body language and posture adjusting clearly to the situation (i.e. open, no guarded posture during an empathetic exchange followed by subtle, no-obvious movement to a more tactical posture as a threat emerges)

7. Self-Assessment – Self-Assessment is scored after the scenario during the debrief when the participant discusses what occurred and their assessment of how they feel about their performance. Additionally, the self-assessment also includes the level of detail the participant is able to articulate regarding what occurred as an indication of the connection between their executive function and recall of sensory input.

- 0 – Participant's self-assessment is clearly out of sync with what happened in the scenario (i.e. if participant shoots an inappropriate target but still asserts that it was a good outcome or the participant does little to impact the scenario other than shooting the immediate threat and assesses they performed very well)
- 1 – Participant's self-assessment is not well-aligned with actual performance or participant is not able to provide any magnifying details of what occurred
- 2 – Participant's self-assessment is somewhat aligned with actual performance and provides some illuminating details to what occurred
- 3 – Participant's self-assessment is aligned with observed performance and provides good details of the event
- 4 – Participant's self-assessment is aligned with performance as well as provides thoughtful insight into their experience coupled with great, accurate detail of the events as they occurred

8. Rationality – Rationality is scored after the scenario during the debrief when the participant discusses what occurred. This criterion refers to the validity of the participant's rationale for their decision making during the scenario in terms of supporting their assessment of how they were able to intervene. For example, not just the decision to shoot a role player that presented a threat but why and when they made that decision.

- 0 – Participant provides no real rationale for their actions or their rationale is illogical or based in inaccurate assessment of the environment
- 1 – Participant only provides rudimentary supporting rationale for their actions (i.e. Simply says, "I shot them because I saw a gun." With no other supporting statements)
- 2 – Participant provides some logical rationale for their decisions but little additional or compelling support
- 3 – Participant provides logical rationale with valid support
- 4 – Participant provides uniquely valid rationale with strong supporting statements for their decision making (i.e. I saw the threat presented but knew that there was still an opportunity to intervene without using lethal force, so I waited to take the shot in order to continue to my strategy for de-escalation, knowing that I still had a tactical advantage if I needed to use lethal force)

References

1. Jackson, A.F.; Bolger, D.J. The neurophysiological bases of EEG and EEG measurement: A review for the rest of us. *Psychophysiology* **2014**, *51*, 1061–1071. https://doi.org/10.1111/psyp.12283.
2. Kübler, A.; Dixon, V.; Garavan, H. Automaticity and reestablishment of executive control-an fMRI study. *Journal of cognitive neuroscience* **2006**, *18*, 1331–1342. https://doi.org/10.1162/JOCN.2006.18.8.1331.
3. Goel, V.; Grafman, J.; Sadato, N.; Hallett, M. Modeling other minds. *Neuroreport* **1995**, *6*, 1741–1746. https://doi.org/10.1097/00001756-199509000-00009.
4. Gallagher, H.L.; Jack, A.I.; Roepstorff, A.; Frith, C.D. Imaging the Intentional Stance in a Competitive Game. *NeuroImage* **2002**, *16*, 814–821. https://doi.org/10.1006/NIMG.2002.1117.
5. Zhang, J.X.; Leung, H.C.; Johnson, M.K. Frontal activations associated with accessing and evaluating information in working memory: an fMRI study. *NeuroImage* **2003**, *20*, 1531–1539. https://doi.org/10.1016/J.NEUROIMAGE.2003.07.016.
6. Pochon, J.B.; Levy, R.; Fossati, P.; Lehericy, S.; Poline, J.B.; Pillon, B.; Bihan, D.L.; Dubois, B. The neural system that bridges reward and cognition in humans: An fMRI study. *Proceedings of the National Academy of Sciences of the United States of America* **2002**, *99*, 5669. https://doi.org/10.1073/PNAS.082111099.
7. Raye, C.L.; Johnson, M.K.; Mitchell, K.J.; Reeder, J.A.; Greene, E.J. Neuroimaging a Single Thought: Dorsolateral PFC Activity Associated with Refreshing Just-Activated Information. *NeuroImage* **2002**, *15*, 447–453. https://doi.org/10.1006/NIMG.2001.0983.
8. Nakai, T.; Kato, C.; Matsuo, K. An fMRI Study to Investigate Auditory Attention: A Model of the Cocktail Party Phenomenon. *Magnetic Resonance in Medical Sciences* **2005**, *4*, 75–82. https://doi.org/10.2463/MRMS.4.75.
9. Fincham, J.M.; Carter, C.S.; Veen, V.V.; Stenger, V.A.; Anderson, J.R. Neural mechanisms of planning: A computational analysis using event-related fMRI. *Proceedings of the National Academy of Sciences of the United States of America* **2002**, *99*, 3346–3351. https://doi.org/10.1073/PNAS.052703399/ASSET/12213FB1-4645-47DD-98CE-C5CA6BCCCD07/ASSETS/GRAPHIC/PQ0527033006.JPEG.
10. Bermpohl, F.; Pascual-Leone, A.; Amedi, A.; Merabet, L.B.; Fregni, F.; Gaab, N.; Alsop, D.; Schlaug, G.; Northoff, G. Attentional modulation of emotional stimulus processing: An fMRI study using emotional expectancy. *Human Brain Mapping* **2006**, *27*, 662. https://doi.org/10.1002/HBM.20209.
11. Seghier, M.L. The angular gyrus: Multiple functions and multiple subdivisions. *Neuroscientist* **2013**, *19*, 43–61. https://doi.org/10.1177/1073858412440596.

sensors

MDPI

Article

Autonomous Learning of New Environments with a Robotic Team Employing Hyper-Spectral Remote Sensing, Comprehensive In-Situ Sensing and Machine Learning

David J. Lary *, David Schaefer, John Waczak, Adam Aker, Aaron Barbosa, Lakitha O. H. Wijeratne, Shawhin Talebi, Bharana Fernando, John Sadler, Tatiana Lary and Matthew D. Lary

Hanson Center for Space Sciences, University of Texas at Dallas, Richardson, TX 75080, USA; captdaveschaefer@gmail.com (D.S.); John.Waczak@utdallas.edu (J.W.); Adam.Aker@utdallas.edu (A.A.); aaronbarbosa.me@gmail.com (A.B.); lhw150030@utdallas.edu (L.O.H.W.); Shawhin.Talebi@utdallas.edu (S.T.); ashen.fernando@utdallas.edu (B.F.); jcs170001@utdallas.edu (J.S.); txl160130@utdallas.edu (T.L.); mdlary@me.com (M.D.L.)

* Correspondence: David.Lary@utdallas.edu

Abstract: This paper describes and demonstrates an autonomous robotic team that can rapidly learn the characteristics of environments that it has never seen before. The flexible paradigm is easily scalable to multi-robot, multi-sensor autonomous teams, and it is relevant to satellite calibration/validation and the creation of new remote sensing data products. A case study is described for the rapid characterisation of the aquatic environment, over a period of just a few minutes we acquired thousands of training data points. This training data allowed for our machine learning algorithms to rapidly learn by example and provide wide area maps of the composition of the environment. Along side these larger autonomous robots two smaller robots that can be deployed by a single individual were also deployed (a walking robot and a robotic hover-board), observing significant small scale spatial variability.

Keywords: machine learning; hyper-spectral imaging; robot team; autonomous; UAV; robotic boat

Citation: Lary, D.J.; Schaefer, D.; Waczak, J.; Aker, A.; Barbosa, A.; Wijeratne, L.O.H.; Talebi, S.; Fernando, B.; Sadler, J.; Lary, T.; et al. Autonomous Learning of New Environments with a Robotic Team Employing Hyper-Spectral Remote Sensing, Comprehensive In-Situ Sensing and Machine Learning. *Sensors* **2021**, *21*, 2240. https://doi.org/10.3390/s21062240

Academic Editor: Mehmet Yuce

Received: 5 January 2021
Accepted: 5 March 2021
Published: 23 March 2021

1. Introduction

This paper describes a robotic team that can rapidly learn new environments. The system that is described here demonstrates a flexible paradigm that is easily scalable to multi-robot, multi-sensor autonomous teams.

The inspiration for this autonomous robotic is the automation of what is currently done manually in the production of remote sensing satellite data products. The typical timescale from starting work on a new remote sensing data product to its operational readiness is at least a couple of years, but, more typically, a decade or more. A key part of this substantial time delay is due to the time that is taken for the collection of the relevant training data. Hence, our goal was to reduce this timescale to be near real time by utilising an autonomous robotic team that can both collect the training data, and then in real time process and stream the remote sensing data products.

A case study is described in detail for the rapid characterisation of the aquatic environment. Other authors have described, in detail, various configurations of autonomous robots, for example [1–6]. Here, we leverage our past experience over the last two decades in pioneering the use of machine learning for providing and calibrating remote sensing data products [7–24] and use it to inform the design and operation of the robotic team.

The aquatic environment was chosen, as it includes extra challenges with regards the ease of access, further demonstrating the value of the approach. When considering the usefulness of being able to conduct such rapid surveys, it is worth noting that, for just the oil spill response use case alone, the National Academy of Sciences estimates that the annual oil spill quantities range from 1.7 million tons to 8.8 million tons. Over 70% of this release

is due to human activities. The result of these spills include dead wildlife, contaminated water, and oil-covered marshlands [25–28]. Accordingly, being able to rapidly survey such areas to guide clean-up operations is of considerable use. It is also of use in a wide variety of contexts, from general environmental surveys, to studying harmful algal blooms, to the clean-up operations after natural disasters, such as hurricanes, etc.

In the example that is described in this paper, the fully autonomous team includes a robotic boat that carries a suite of sensors to measure water composition in real time as well as a sonar, and an autonomous UAV equipped with a down-welling irradiance spectrometer, hyper-spectral, and thermal imagers, together with an onboard Machine Learning (ML) capability. Figure 1 shows photographs of the robot team during a December 2020 deployment in North Texas.

Figure 1. Photographs of the robot team during a Fall 2020 deployment in North Texas.

Besides this capability being useful by itself, there is a wider significance for earth observing satellite missions. A key component to each and every space agency earth observation mission is the delivery of a suite of data products and the calibration/validation of these products. The demonstrated paradigm can reduce the time and cost of producing new remote sensing data products, while increasing the functionality and data quality and providing new real-time automated calibration/validation capabilities.

The approach also provides enhanced capabilities for real-time onboard data product creation, reducing product delivery latency. The end-to-end demonstration uses all off-the-shelf components, representing a reduction in costs and risk when prototyping new mission concepts.The use of embedded machine learning is a key element, so we will refer to the approach as Rapid Embedded Prototyping for Advanced Applications (REPAA).

Hyper-Spectral Imaging

The human eye perceives the color of visible light in three bands using the cones, the photoreceptor cells in the retina (Figure 2). These three broad bands are red (centered on 564 nm), green (centered on 534 nm), and blue (centered on 420 nm). By contrast, instead of using just three broad bands, hyper-spectral cameras divide the spectrum into a very large number of narrow bands, in our case 463 bands from 391–1011 nm. A hyper-cube is a three-dimensional dataset that consists of a stack of two-dimensional image layers each for a different wavelength. Hence, for each pixel in the image, we have a multi-wavelength spectra (spectral signature). This is schematically shown in the lower left of Figure 2. On the right, we see a conventional RGB color image with only three bands, images for red, green, and blue wavelengths.

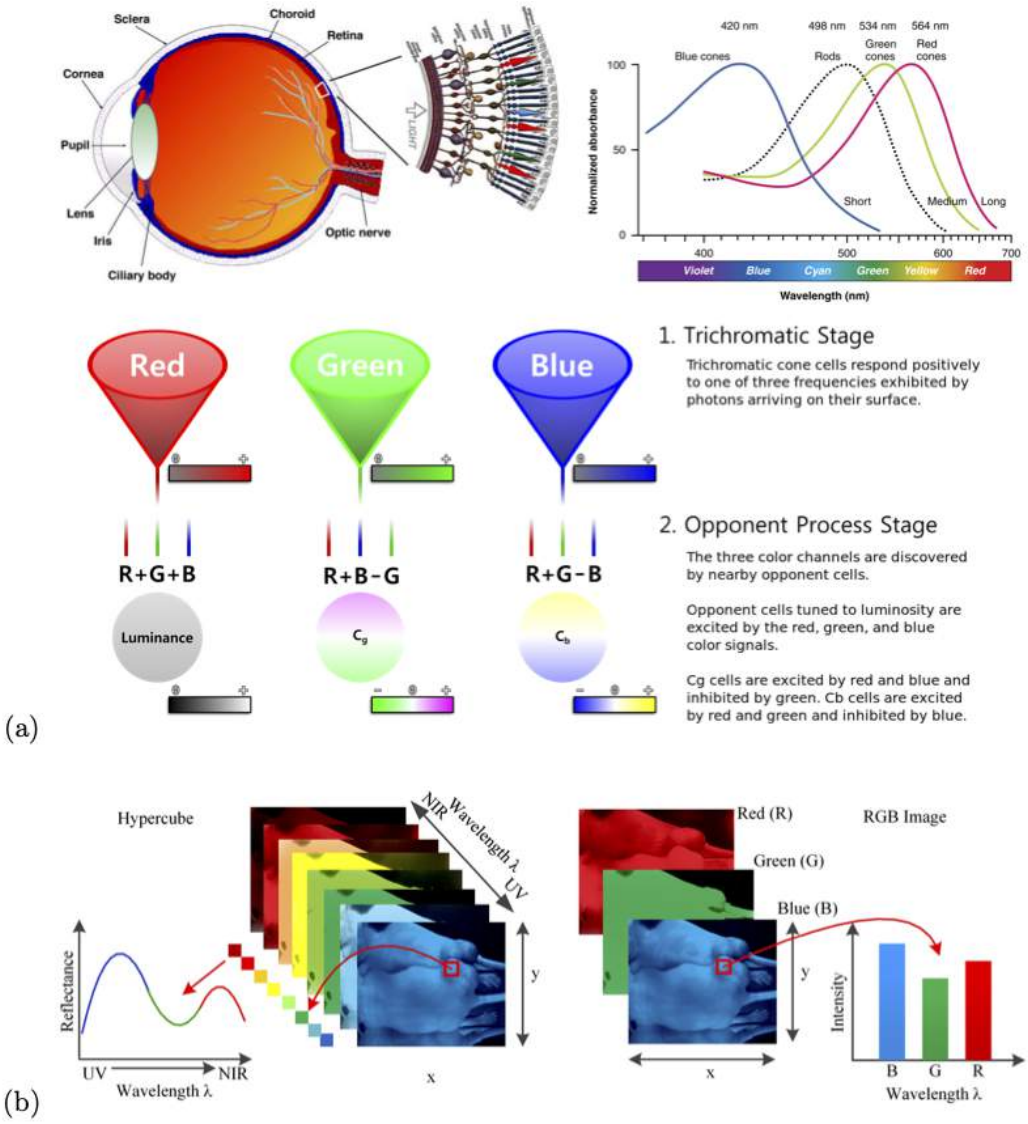


Figure 2. Panel (**a**) Trichromatic cone cells in the eye respond to one of three wavelength ranges (RGB). Panel (**b**) shows a comparison between a hyper-spectral data-cube and RGB images.

Chemicals absorb light in a characteristic way. Their absorption spectra is a function of their chemical structure. Figure 3a shows the structure of chlorophyll and the associated absorption spectra. So that we can accurately calculate the reflectivity at each wavelength our autonomous UAV measures both the incident downwelling irradiance of incident solar radiation and a hyper-spectral imager pointed directly down at the earth's surface below the UAV. For every pixel we measure an entire spectrum with a hyper-spectral camera so we can identify chemicals within the scene.

Figure 3b shows an example reflectivity hyper-spectral data cube collected during a robot team deployment in North Texas during November 2020. This data cube includes the area where an inert dye was released to test the system. The dye used was Rhodamine WT, a fluorescent, xanthene dye, which has long been used as a hydrologic tracer in surface water systems. The spectral signature of the dye is clearly visible in the hyper-spectral data cube. The top layer of the hyper-spectral data cube shows the regular RGB image, the 463 stacked layers below show the reflectivity (on a log-scale) for each wavelength band between 391 and 1011 nm.

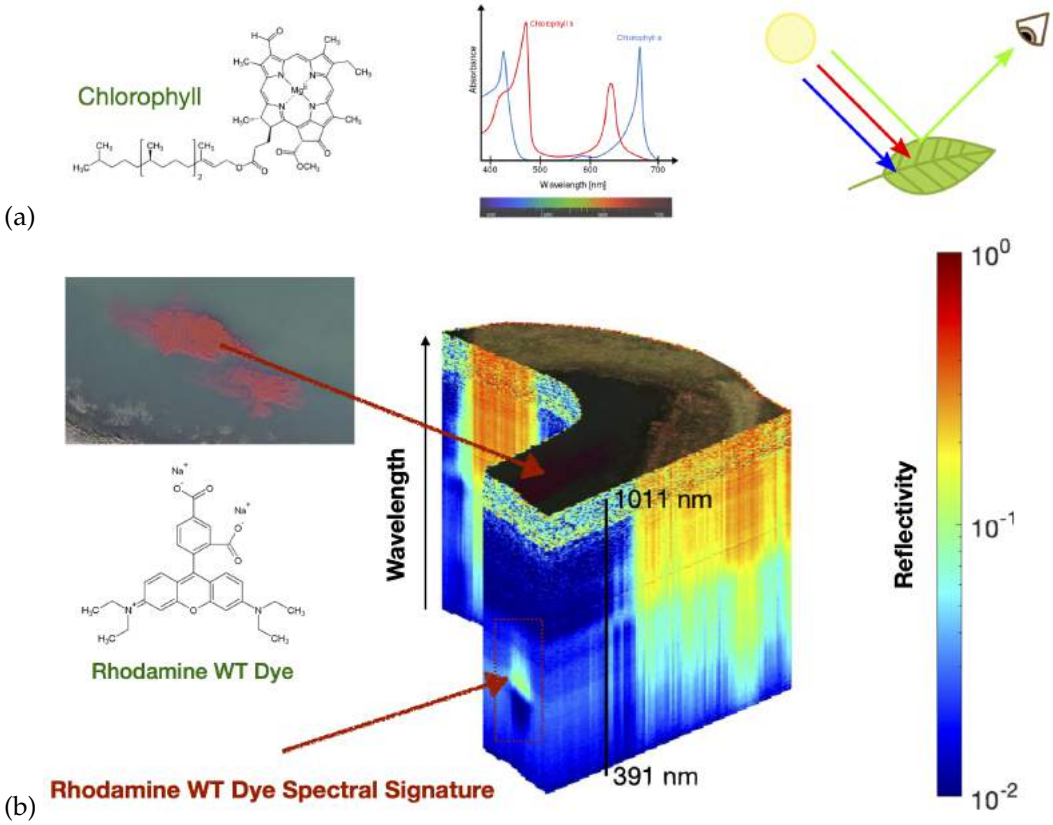


Figure 3. Panel (**a**) Chemicals absorb light in a characteristic way. Their absorption spectra is a function of their chemical structure. For every pixel we measure an entire spectrum with a hyper-spectral camera so we can identify chemicals within the scene. Panel (**b**) shows an example hyper-spectral data cube collected in North Texas on 23 November 2020. This particular data cube includes a simulant release, Rhodamine WT. The top layer of the hyper-spectral data cube shows the regular RGB image, the 462 stacked layers below show the reflectivity (on a log-scale) for each wavelength band between 391 and 1011 nm.

2. Materials and Methods

All of the data for the machine learning data product creation were collected in a coordinated automated manner using the autonomous robotic team.

The most time consuming process in building this robotic team was finding the appropriate off the shelf components to implement the prototype. So here we have provided the full detailed recipe on the autonomous robotic team in the hope that it will facilitate the research of others.

An overview of the robotic team members and their sensor payloads is as follows.

2.1. Robotic Vehicles

A Maritime Robotics Otter (https://www.maritimerobotics.com/otter, accessed 5 January 2021) autonomous boat was used. With a footprint of only 200 × 108 × 81.5 cm, a weight of 55 kg, and dual electrical fixed thrusters, it is an easily deployable asset that can be transported in a van or even within normal airliners to a survey site. With a cruise speed of two knots, it has a duration of 20 h from one charge of the batteries. It can use WiFi, cellular, and an optional AIS receiver for communication to the control station.

A Freefly Alta-X (https://freeflysystems.com/alta-x, accessed 5 January 2021) autonomous professional quad-copter was used. It was specifically designed to carry cameras, with a payload capacity of up to 35 lb, a long range data link, and autonomy provided by the Open PX4 flight stack. The open source QGroundControl software was used to control the autonomous operations (https://freeflysystems.com/support/alta-pro-support, accessed 5 January 2021). QGroundControl is available for Mac, Windows, iOS, and Android.

All of the robotic team members carry a high-accuracy GPS and INS, so that every data point can be geo-located and time stamped. Each of the robots can also join the same network which connects the robots and their ground-control stations. Our robots use long-range Ubiquiti 5 GHz LiteBeam airMAX WiFi (https://www.ui.com, accessed 5 January 2021). The airMAX Time Division Multiple Access (TDMA) protocol allows for each client to send and receive data using pre-designated time slots that are managed by an intelligent AP controller. This time slot method eliminates hidden node collisions and maximizes airtime efficiency. This WiFi network is connected to the internet using a Cradlepoint cellular modem (https://cradlepoint.com, accessed 5 January 2021).

This network also includes a local Synology network-attached storage (NAS) (https://www.synology.com, accessed 5 January 2021) device in the robot team control trailer, which, in real-time, syncs the data that were collected to the NAS in our home laboratory in the university.

2.2. Boat Sensors

The robotic boat payload included a BioSonics MX Aquatic Habitat Echosounder sonar for rapid assessment and mapping of aquatic vegetation, substrate and bathymetry (https://www.biosonicsinc.com/products/mx-aquatic-habitat-echosounder/, accessed 5 January 2021). Three Eureka Manta-40 multi-probes (https://www.waterprobes.com/multiprobes-and-sondes-for-monitori, accessed 5 January 2021), a Sequoia Scientific LISST-ABS acoustic backscatter sediment sensor (https://www.sequoiasci.com/product/lisst-abs/, accessed 5 January 2021), and an Airmar Technology Corporation 220 WX ultra-sonic weather monitoring sensor (https://www.airmar.com/weather-description.html?id=153, accessed 5 January 2021).

The first Manta-40 multi-probe included sensors for temperature and turbidity and Turner Designs Cyclops-7 submersible Titanium body fluorometers (https://www.turnerdesigns.com/cyclops-7f-submersible-fluorometer, accessed 5 January 2021) for Chlorophyll A, Chlorophyll A with Red Excitation, Blue-Green Algae for fresh water (Phycocyanin), Blue-Green Algae for salt water (Phycoerythrin), and CDOM/FDOM. The second Manta-40 multi-probe included sensors for temperature, conductivity (with specific conductance, salinity, and total dissolved solids, TDS), pH (with separate reference electrode), optical dissolved-oxygen, turbidity, and Ion Selective Electrodes by Analytical Sensors and Instruments (http://www.asi-sensors.com/, accessed 5 January 2021) for ammonium (NH_4^+), bromide (Br^-), calcium (Ca^{++}), chloride (Cl^-), nitrate (NO_3^-), and sodium (Na^+). The third Manta-40 multi-probe included sensors for temperature, turbidity, a total dissolved gas sensor, and Turner Designs Cyclops-7 submersible Titanium body fluorometers for optical brighteners, crude oil, refined fuels, and tryptophan.

In addition, a portable Membrane Inlet Mass Spectrometer (MIMS) designed and built by Prof. Verbeck of the University of North Texas is available (but not used in these deployments) to switch every 3 s between sampling the water composition and the air composition.

2.3. Aerial Sensors

The aerial vehicle used a Gremsy H16 gimbal (https://gremsy.com/gremsy-h16, accessed 5 January 2021) that was made with aircraft grade aluminum and carbon fiber to carry a Resonon Visible+Near-Infrared (VNIR) Pika XC2 (https://resonon.com/Pika-XC2, accessed 5 January 2021) hyper-spectral camera (391–1011 nm) with a Schneider Xenoplan 1.4/17 mm lens, and a FLIR Duo Pro R, (640 $\times$ 512, 25 mm, 30 Hz) combining a high resolution, radiometric thermal imager, 4K color camera, and a full suite of onboard sensors (https://www.flir.com/products/duo-pro-r/, accessed 5 January 2021). On the top of the quad copter there is a sky facing Ocean Optics UV-Vis-NIR spectrometers measuring the incident down-welling irradiance, allowing us to calculate reflectance.

2.4. Geo-Rectification

The hyper-spectral data cubes collected are very large and written in real time to the solid-state disk (SSD) that was attached to the Resonon Pika XC2. The Camera SSD is exported as a Network File System (NFS) mount, so that a second onboard computer can geo-rectify the hyper-spectral data cubes as they are created, in order to facilitate the real-time processing of these files. These hyper-spectral data cubes provide a visible and near infrared spectrum (391–1011 nm) for each pixel. Once these data cubes are geo-rectified in real-time, they are available for onboard machine learning using edge computing onboard the aerial vehicle.

2.5. Machine Learning

The accurate geo-tagging and time stamping of all data from all members of the robot team allows for the automation of the machine learning data product creation. For every location at which the robotic boat sampled the in-situ water composition, we associate a VNIR remotely sensed spectrum (391–1011 nm) that is provided by the hyper-spectral data cubes collected by the aerial-vehicle. This data are then be used for multi-variate non-linear non-parametric machine learning, where the inputs are the spectrum, in this case 462 values from the 391–1011 nm spectra, and the outputs are each of the values measured in-situ by the robotic boat. A variety of machine learning approaches were used. These approaches included shallow neural networks with hyper-parameter optimization, ensembles of hyper-parameter optimized decision trees, gaussian process regression with hyper-parameter optimization, and a super-learner, including all of the previously mentioned approaches. Each empirical non-linear non-parametric fit is evaluated by constructing both a scatter diagram and a quantile-quantile plot of the values estimated by the machine learning model plotted against the actual values in the independent validation dataset.

The use of machine learning in this study builds on our heritage of using machine learning for sensing applications over the last two decades [7–24].

3. Learning Modes

We designed each component of our system to be flexible for different scenarios and deployment configurations. The entire system is called a Cyber Physical Observatory (Figure 4). A few basic definitions/descriptions are helpful in appreciating the benefits of this. The Cyber Physical Observatory is a collection of sentinels and/or robot teams that provide real-time data and actionable insights, and whose capabilities can be updated via an app store. The Robot Team is a collection of co-operative autonomous sentinels. A Sentinel is a Software Defined Sensor that is mounted on a Platform. A Platform supplies the Software Defined Sensor with power, timestamps for all observations, communication, and mobility where applicable. In some of our other applications, these even include wearable sensors. A Software Defined Sensor is a smart sensor package that combines a physical sensing system with software/machine learning, providing a variety of calibrated data products that can be updated via an app store.

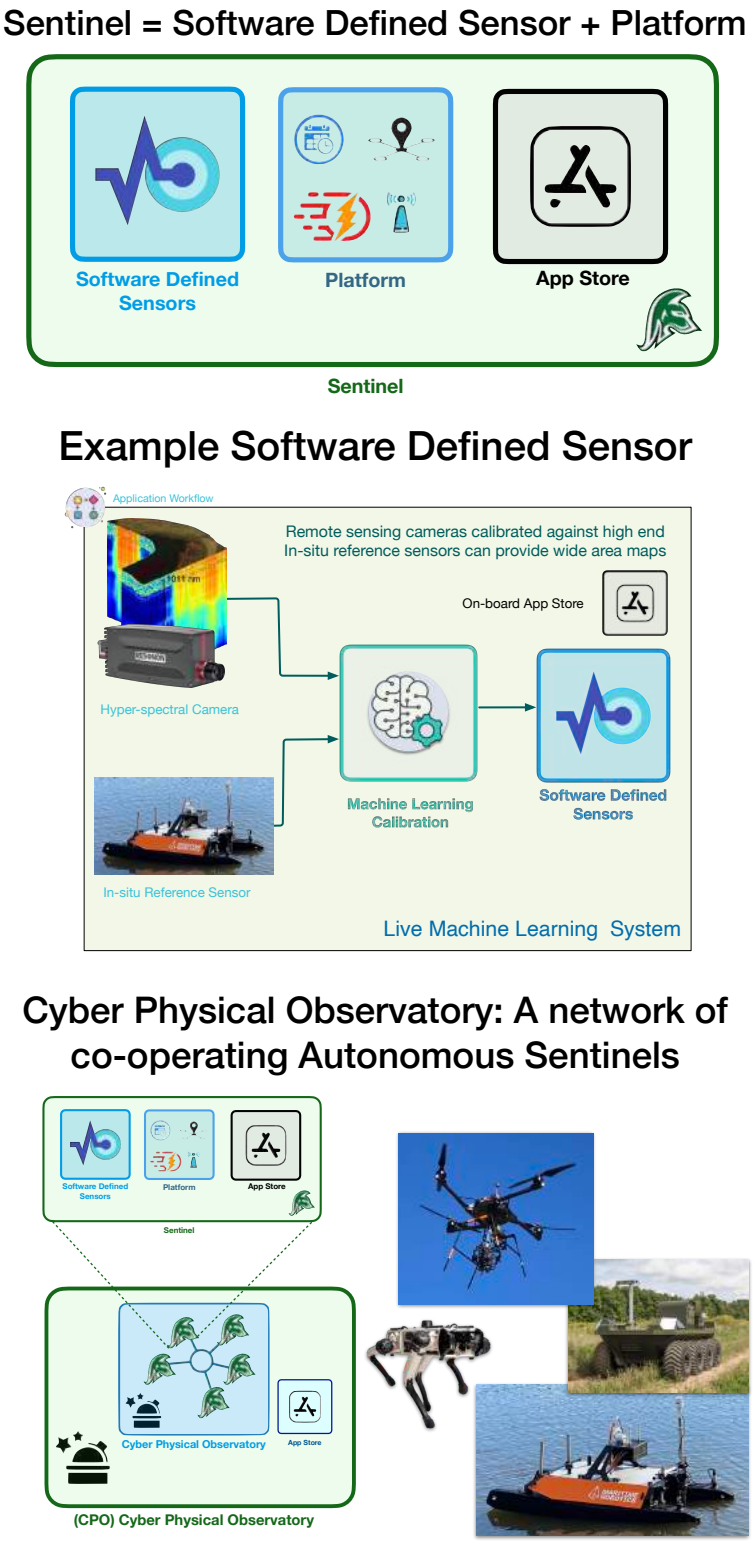


Figure 4. The Cyber Physical Observatory is a collection of sentinels that provide real-time data. A Sentinel is a Software Defined Sensor that is mounted on a Platform. A Platform supplies the Software Defined Sensor with power, timestamps for all observations, communication, and mobility where applicable. A Software Defined Sensor is a smart sensor package that combines a physical sensing system with machine learning providing a variety of calibrated data products that can be updated via an app store.

Two distinct machine learning modalities are useful when trying to rapidly learn new environments (Figure 5). Mode 1: Coordinated robots using onboard Machine Learning for specific data products. Mode 2: Unsupervised classification.

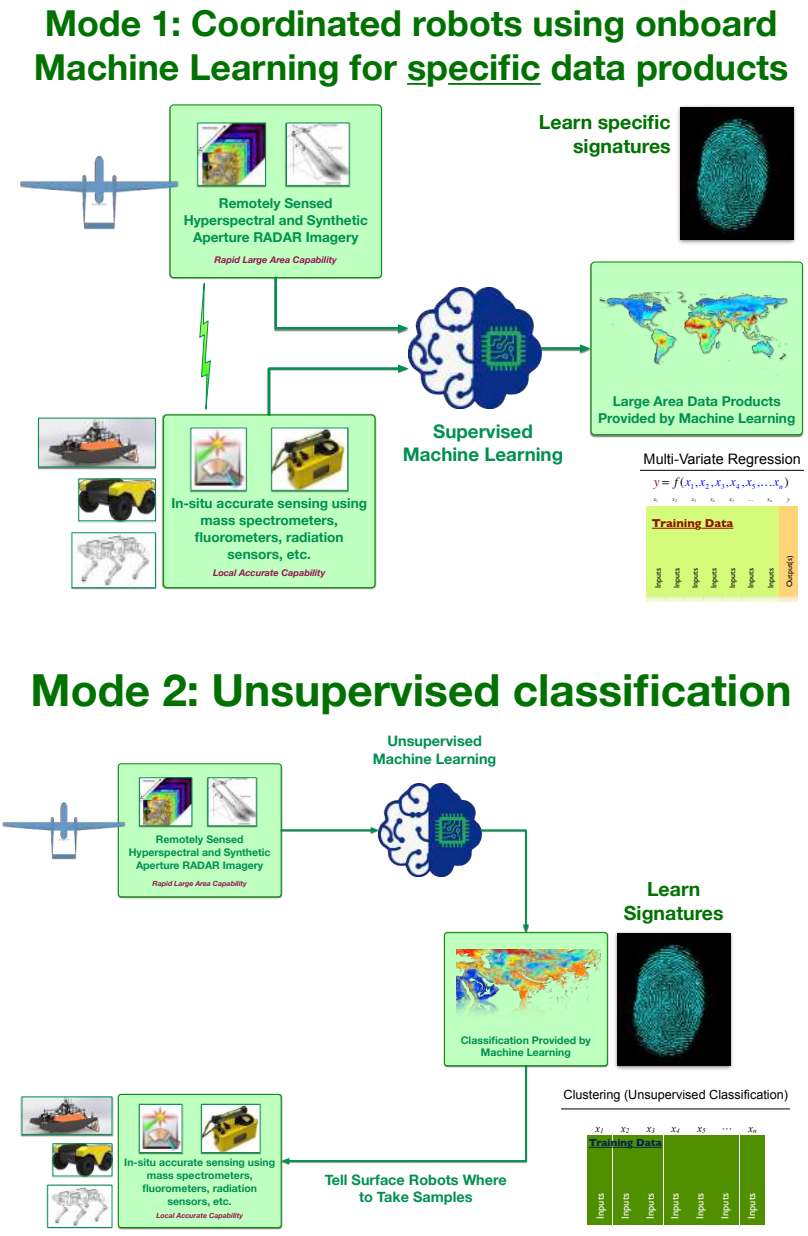


Figure 5. The autonomous robotic team operates in two modes. **Mode 1**: Coordinated robots using onboard Machine Learning for specific data products. **Mode 2**: Unsupervised classification.

In Mode 1, the robot team members rapidly collect the machine learning training data in a carefully coordinated manner. For our example deployment in North Texas during the Fall of 2020, over a period of about fifteen minutes, thousands of precisely collocated measurements were made by the robotic team. The robotic boat autonomously measures in-situ ground truth of a large array of parameters using the sensors described above, while the robotic aerial vehicle gathered remotely sensed observations of exactly the same locations using hyper-spectral and thermal imaging. These remotely sensed observations could be readily extended to cover a wider wavelength range and include Synthetic Aperture Radar (SAR). Once the training data are acquired, the machine learning algorithms can rapidly learn the mapping from the remotely sensed observations to the

in-situ ground truth. Figure 6 shows three different examples of the validation of these autonomously acquired machine learning data products being independently verified while using scatter diagrams and quantile-quantile plots.

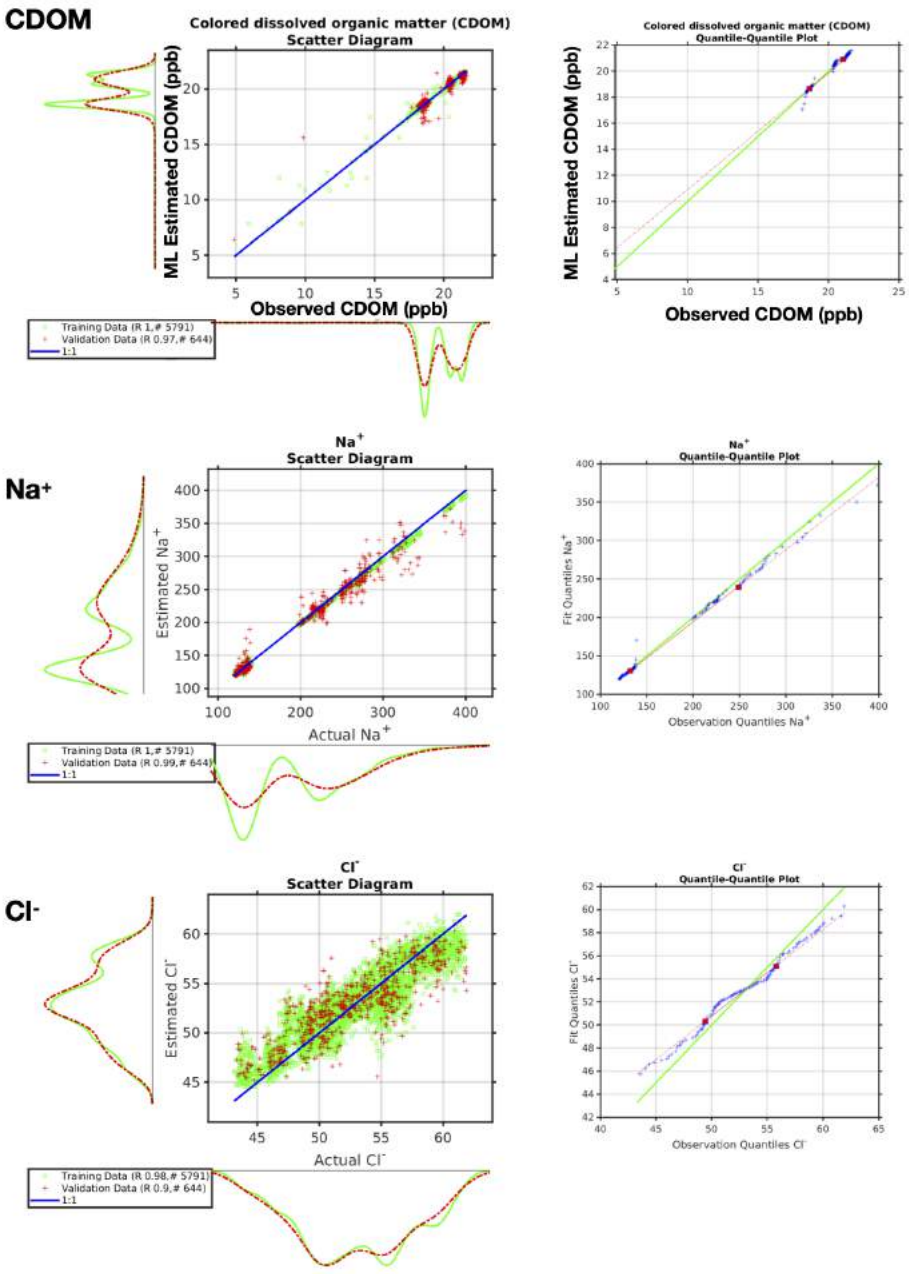


Figure 6. Machine learning performance quantified by both scatter diagrams and quantile-quantile plots utilising data collected autonomously by the robot team during three exercises during November and December 2020 in North Texas. The three examples shown here are for CDOM, Na^+, and Cl^-. The scatter diagrams show the actual observations (mg/L) on the *x*-axis and the machine learning estimate on the *y*-axis. The green curves are for the training data, the red for the independent validation. The legend shows the number of points in the training and validation datasets and their associated correlation coefficients. The quantile-quantile plots show the observation quantiles on the *x*-axis and the machine learning estimate quantiles on the *y*-axis.

Once the machine learning algorithm(s) have been trained, they can then be used to rapidly provide wide-area maps with just the remotely sensed observations. Two examples of this are shown in Figure 7. These can be processed onboard the aerial vehicle

and the results streamed in real-time to the ground control station. The robotic boat can then be autonomously tasked to verify the wide area maps by collecting independent validation data.

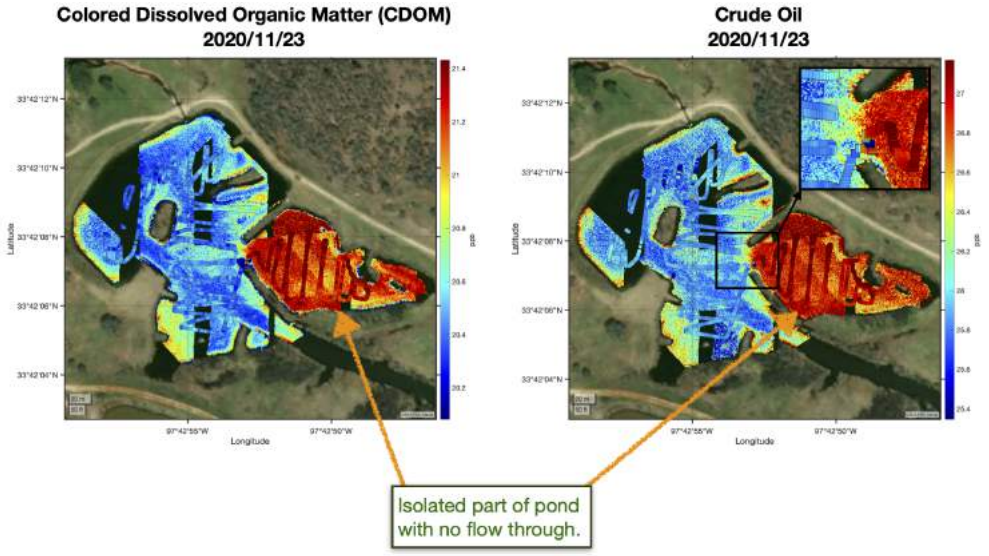


Figure 7. Example crude oil and colored dissolved organic mater (CDOM) data collected autonomously by the robot team on November 23, 2020 in North Texas. The maps show the CDOM and crude oil estimated from the hyper-spectral imager using machine learning as the background colors and the actual in-situ boat observations as the overlaid color filled squares. Note that the isolated part of the pond, which has now fresh water in-flux, has higher levels of CDOM and crude oil with a sharp gradient across the inlet in both the estimates using the hyper-spectral image and the boat observations.

In Mode 2, we would like to perform a fine-grained multi-class surface classification of the entire domain. This is done by providing the remotely sensed data (in this case, the hyper-spectral and thermal imagery) to an unsupervised classification. The unsupervised machine learning characterises the distinct regions and zones in the area of interest. This can be particularly useful when trying to identify the location of particular contaminants, suggesting the optimum sampling patterns that are required beyond the usual clover leaf, star, or box patterns used for contaminant searches.

4. Results

Over a period of just a few minutes, we acquire thousands of training data points. This training data allows for our machine learning algorithms to rapidly learn by example. The machine learning fit used here is an optimized ensemble of regression trees [29–31] with hyper-parameter optimization [32] implemented in Matlab version 2021a (https://www.mathworks.com, accessed 5 January 2021) using the function `fitrensemble` with all hyper-paramter optimization selcted and parallel processing enabled. A loop is executed over all the variables that were measured by the robotic that we would like to estimate using the hyper-spectral imagery.

A balanced training dataset is constructed for each of these variables. This is done by considering each input and output variable in the training dataset in turn and calculating n percentiles, from each of these n percentile ranges covering the entire PDF, from each percentile range we select m random values (where $m < n$) for the training and a different set of random values for independent validation.

Figure 6 shows an example of the colored dissolved organic mater (CDOM) data collected autonomously by the robot team on 23 November 2020 in North Texas, along with some of the aqueous ion data. The panel shows a scatter diagram of the actual observations on the x-axis and the machine learning estimate on the y-axis. The green curves are for

the training data, the red for the independent validation. On each axis, we also show the associated PDFs. The ideal result is shown in blue (a slope of 1 and an intercept of zero for the scatter diagram).

Figure 7 shows maps of the CDOM and crude oil concentration estimated while using the machine learning as the background colors and the actual in-situ boat observations as the overlaid color filled squares. Note that the isolated part of the pond that now has fresh water in-flux has higher levels of CDOM and crude oil with a sharp gradient across the inlet in both of the estimates using the hyper-spectral image and the boat observations. We note that there is good agreement between the machine learning estimate and the actual in-situ boat observations.

5. Discussion

5.1. Limitations

The fidelity of the data products that are provided by the autonomous robotic team is limited by the training data that it is able to acquire. For example, our remote sensing hyper-spectral camera in the demonstration use case presented here observes the spectral region 391–1011 nm. It would be useful to extend this spectral region, so that we can see more chromophores, and to extend the type of remote sensing imaging, e.g., to include Synthetic Aperture RADAR (SAR).

It would also be useful for the boat to have larger pontoons, so that it can carry our mass-spectrometer that can sample both the air and water, switching between the two inlets every three seconds.

We would also like to extend the machine learning approaches to include Physics Based machine learning, such that the machine learning is constrained by known physical principles.

5.2. Automating Data Product Creation

A key factor in providing remotely sensed water composition products is providing a comprehensive database of water composition (e.g., SeaBASS, the publicly shared archive of in-situ oceanographic and atmospheric data maintained by the NASA Ocean Biology Processing Group https://seabass.gsfc.nasa.gov, accessed 5 January 2021). The cost of making the measurements of ocean composition can be substantial, because it involves a significant ship time as well as a large support team. Secondly, because the satellites are in a fixed orbit with a fixed viewing geometry, the number of coincidences between the shipboard water observations and the orbiting satellite observations are, by definition, limited. Typically several thousand coincident observations are used in the tuning and creation of a NASA ocean data product. In the REPAA approach, the entire system can be automated and objectively optimized. Thus, with a data rate of one observation every second, in a matter of hours we can gather tens of thousands of observations in a totally automated, fully coordinated manner, as was demonstrated in North Texas during November and December 2020 (Figure 1). There is explicit coordination between the water observations that were taken from the robotic boat and the continuous aerial observations made by the robotic aerial vehicle carrying a hyper-spectral imager. The system can be deployed to very diverse environments across a matter of just weeks to months, so, over a matter of just weeks to months, millions of coordinated, precisely coincident records can be made. Furthermore, we have previously demonstrated, the data can be randomly partitioned into training and independent validation sets, and using the onboard machine learning, transformed into optimal water composition data products, using many orders of magnitude more observations than before at a fraction of the cost and in a fraction of the time.

Aurin et al. [33] provides one of the most comprehensive training datasets to date for Chromophoric Dissolved Organic Matter (CDOM). Their Global Ocean Carbon Algorithm Database (GOCAD) for Chromophoric Dissolved Organic Matter (CDOM) encompasses 20,000–100,000+ records (depending on the variable considered) and it is based on oceano-

graphic campaigns that were conducted across the world over the past 30 years at great expense. In contrast, the autonomous robotic team can collect around 20,000+ precisely coordinated training records per hour. By design, the robotic team makes precisely coordinated overpasses of exactly the same locations; this leads to providing a training dataset with a high data rate. By deploying the team on multiple occasions at a diversity of locations, one can rapidly build a comprehensive training dataset.

The traditional approach for creating remote sensing data products, as shown on the left of Figure 8, is compared with the approach that was used in this study, as shown on the right. Using the REPAA approach, data collection and the creation of derivative data products can be carried out on the same day, for example, in the December 2020 exercises in North Texas (Figure 1).

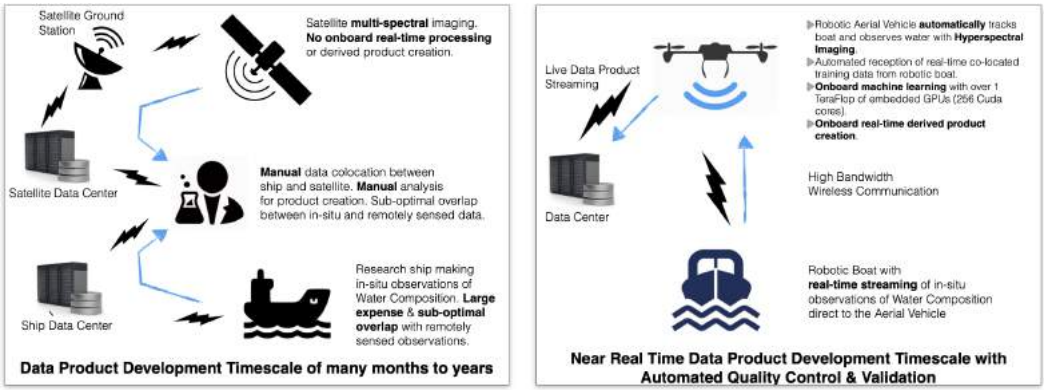


Figure 8. Schematics illustrating the traditional approach to creating remote sensing data products (**left**) and that used in this study (**right**).

5.3. Improving Product Quality & Automating Cal/Val

Critical in improving product quality is the comprehensive training data set, which spans as much parameter space and variability that is actually found in the real world. This necessitates making observations in a large number of diverse contexts. Being able to make these observations with such a highly automated platform is a tremendous step forward and it costs less. In summary, our robotic platform can address the issue of small scale variability encountered across a satellite pixel. These capabilities assist in continuing validation/quality control and it can help to optimize the waveband selection for future satellite instruments and missions.

5.4. Reducing Latency for Product Delivery as Well as Mission Risk, Cost, Weight and Size

Utilizing new embedded onboard processing (1 TeraFlop weighing just 88 g with a size of only 87 mm $\times$ 50 mm) for real-time onboard processing leads to reducing the latency in product delivery from hours/days to just the downlink time. The product delivery latency can be critical for decision support applications, such as oil spills, or other disaster response applications, and for routine forecasting and data assimilation applications. A risk reduction is also realized, by the ability to first deploy an end-to-end demonstrator, while using entirely commercial off the shelf components and low cost aerial vehicles, with all software being made Open Source.

5.5. Onboard App Store

There is currently a rapid enhancement in both observing capabilities and the embedded computing power from miniaturized low power devices. As these enhanced observing capabilities become routinely available on small cubesats (like hyperspectral imaging), the number of possible uses and applications for societal benefit grows. However, so does the bandwidth that is required for the downlink of the hyperspectral datacubes. Hence, the possibility of onboard processing, for example, using embedded machine learning, means that product creation can occur directly onboard the cubesats and then streamed

live via the downlink. This reduces the latency of product creation and the bandwidth that is needed for the downlink. The next logical step, then, of a rapid prototyping and agile workflow, is an onboard app store, where new data products can be deployed to the remote sensing platform for seamless use onboard. A formalized development, testing, and deployment workflow with an app store facilitates an Earth-observing system that responds to the rapidly changing societal needs while maintaining a rigorous approach to validation. This onboard app store can leverage the smart automated code generation that already exists off the shelf and is now routinely used for automobiles and aircraft across the world. The time has also come for this to be the standard paradigm for earth observation.

5.6. Smaller Robots

There is also value in smaller robots that are easy to transport by a single individual. Figure 9 shows photographs of the smaller walking robot (from Ghost Robotics) and a robotic hover-board (conceived and built by Aaron Barbosa) that we deployed along size the larger autonomous robotic team for illustrative purposes. The walking robot and robotic hover-board both carried exactly the same payload of sensors that could be rapidly switched between the robots. The sensing payload measured, every few seconds, the full size spectrum of airborne particulates in the size range 0.3–43 microns and the abundance of a selection of gases. The laser scanner onboad the walking robot acquired a map of the vicinity, while also measuring in-situ the atmospheric composition, finding very localized changes in the abundance of the airborne particulates of various sizes.

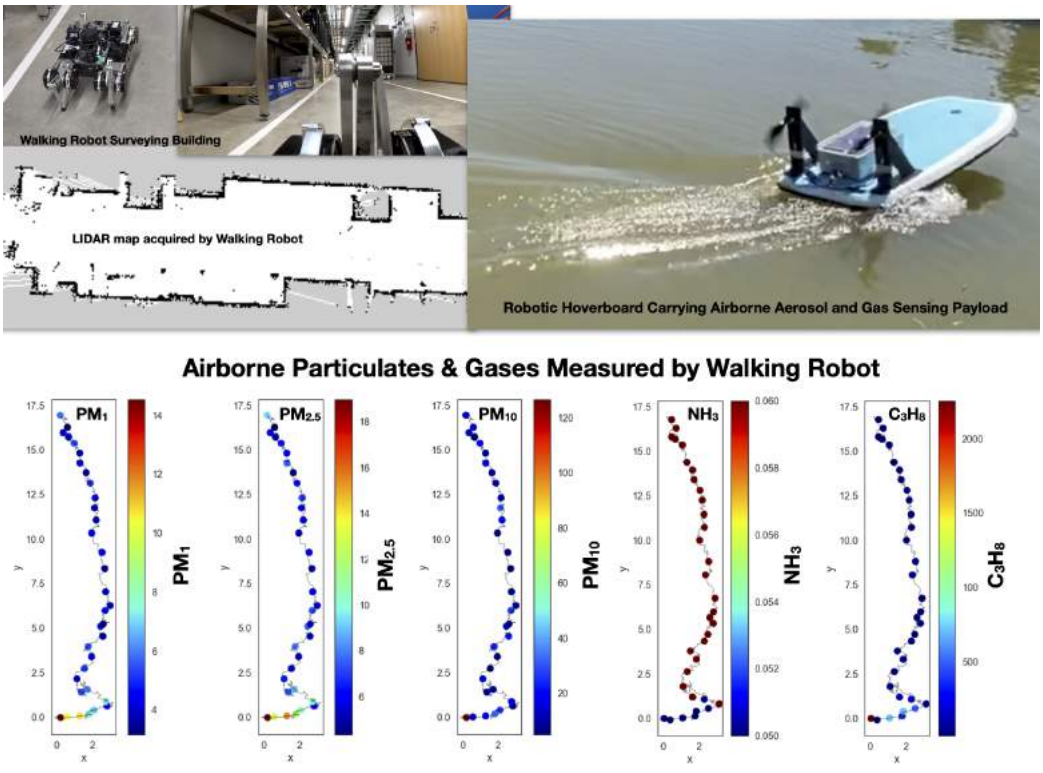


Figure 9. Photographs of the smaller walking robot (from Ghost Robotics) and a robotic hover-board (conceived and built by Aaron Barbosa). For illustrative purposes both of these small robots carried exactly the same payload of sensors measuring the size spectrum of airborne particulates in the size range 0.3–43 microns and the abundance of a selection of gases. The laser scanner onboad the walking robot acquired a map of the vicinity, while also measuring in-situ the atmospheric composition, finding very localized changes in the abundance of the airborne particulates of various sizes.

6. Conclusions

This paper described and demonstrated an autonomous robotic team that can rapidly learn the characteristics of environments that it has never seen before. The flexible paradigm is easily scalable to multi-robot, multi-sensor autonomous teams, and it is relevant to satellite calibration/validation and the creation of new remote sensing data products. A case study was described for the rapid characterisation of the aquatic environment; over a period of just a few minutes, we acquired thousands of training data points. This training data allowed our machine learning algorithms to rapidly learn by example and provide wide area maps of the composition of the environment. Alongside these larger autonomous robots, two smaller robots that can be deployed by a single individual were also deployed, a walking robot and a robotic hover-board, each measuring the full size spectrum of airborne particulates in the size range of 0.3–43 microns and a selection of gases. Significant small scale spatial variability was evident in these hyper-localized observations.

Author Contributions: Conceptualization, D.J.L.; methodology, D.J.L.; software, D.J.L. and J.W.; field deployment and preparation D.J.L., D.S., J.W., A.A., A.B., L.O.H.W., S.T., B.F., J.S., M.D.L., and T.L.; validation, D.J.L.; formal analysis, D.J.L.; investigation, D.J.L.; resources, D.J.L.; data curation, D.J.L., J.W., A.A. and L.O.H.W.; writing—original draft preparation, D.J.L.; writing—review and editing, D.J.L.; visualization, D.J.L.; supervision, D.J.L.; project administration, D.J.L.; funding acquisition, D.J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the following grants: The Texas National Security Network Excellence Fund award for Environmental Sensing Security Sentinels. SOFWERX award for Machine Learning for Robotic Team. Support from the University of Texas at Dallas Office of Sponsored Programs, Dean of Natural Sciences and Mathematics, and Chair of the Physics Department are gratefully acknowledged. The authors acknowledge the OIT-Cyberinfrastructure Research Computing group at the University of Texas at Dallas and the TRECIS CC* Cyberteam (NSF 2019135) for providing HPC resources that contributed to this research (https://utdallas.edu/oit/departments/circ/, accessed 5 January 2021).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Don MacLaughlin, Scotty MacLaughlin and the City of Plano, TX, are gratefully acknowledged for allowing us to deploy the autonomous robot team on their land. Without their assistance this work would not have been possible. Christopher Simmons is gratefully acknowledged for his computational support. Annette Rogers is gratefully acknowledged for support with arranging insurance coverage. Steven Lyles is gratefully acknowledged for support arranging a secure place for the robot team. The authors acknowledge the OIT-Cyberinfrastructure Research Computing group at the University of Texas at Dallas and the TRECIS CC* Cyberteam (NSF #2019135) for providing HPC resources that contributed to this research (https://utdallas.edu/oit/departments/circ/, accessed 5 January 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CDOM	Chromophoric Dissolved Organic Matter
GOCAD	Global Ocean Carbon Algorithm Database
GPS	Global Positioning System
INS	Inertial Navigation System
MIMS	Membrane Inlet Mass Spectrometer
ML	Machine Learning
NASA	The National Aeronautics and Space Administration
NFS	Network File System

REPAA	Rapid Embedded Prototyping for Advanced Applications
SeaBASS	SeaWiFS Bio-optical Archive and Storage System
SSD	Solid State Disk
UAV	Unmanned Aerial Vehicle
VNIR	Visible and Near-Infrared

References

1. Ingrand, F.; Ghallab, M. Deliberation for autonomous robots: A survey. *Artif. Intell.* **2017**, *247*, 10–44. [CrossRef]
2. Islam, M.J.; Hong, J.; Sattar, J. Person-following by autonomous robots: A categorical overview. *Int. J. Robot. Res.* **2019**, *38*, 1581–1618. [CrossRef]
3. Strom, D.P.; Bogoslavskyi, I.; Stachniss, C. Robust exploration and homing for autonomous robots. *Robot. Auton. Syst.* **2017**, *90*, 125–135. [CrossRef]
4. Levent, H. Managing an Autonomous Robot Team: The Cerberus Team Case Study. *Int. J. Hum. Friendly Welf. Robot. Syst.* **2005**, *6*, 35–40
5. Davis, J.D.; Sevimli, Y.; Ackerman, M.; Chirikjian, G. A Robot Capable of Autonomous Robotic Team Repair: The Hex-DMR II System. In *Advances in Reconfigurable Mechanisms and Robots II*; Springer: Cham, Switzerland, 2016.
6. Husain, A.; Jones, H.; Kannan, B.; Wong, U.; Pimentel, T.; Tang, S.; Daftry, S.; Huber, S.; Whittaker, W. Mapping planetary caves with an autonomous, heterogeneous robot team. In Proceedings of the 2013 IEEE Aerospace Conference, Big Sky, MT, USA, 2–9 March 2013; pp. 1–13.
7. Brown, M.E.; Lary, D.J.; Vrieling, A.; Stathakis, D.; Mussa, H. Neural networks as a tool for constructing continuous NDVI time series from AVHRR and MODIS. *Int. J. Remote. Sens.* **2008**, *29*, 7141–7158. [CrossRef]
8. Lary, D.; Aulov, O. Space-based measurements of HCl: Intercomparison and historical context. *J. Geophys. Res. Atmos.* **2008**, *113*. [CrossRef]
9. Lary, D.J.; Remer, L.; MacNeill, D.; Roscoe, B.; Paradise, S. Machine learning and bias correction of MODIS aerosol optical depth. *IEEE Geosci. Remote. Sens. Lett.* **2009**, *6*, 694–698. [CrossRef]
10. Lary, D.; Waugh, D.; Douglass, A.; Stolarski, R.; Newman, P.; Mussa, H. Variations in stratospheric inorganic chlorine between 1991 and 2006. *Geophys. Res. Lett.* **2007**, *34*. [CrossRef]
11. Lary, D.; M′uller, M.; Mussa, H. Using neural networks to describe tracer correlations. *Atmos. Chem. Phys.* **2004**, *4*, 143–146. [CrossRef]
12. Lary, D.J. *Artificial Intelligence in Geoscience and Remote Sensing*; INTECH Open Access Publisher: London, UK, 2010.
13. Malakar, N.K.; Knuth, K.H.; Lary, D.J. Maximum Joint Entropy and Information-Based Collaboration of Automated Learning Machines. In *AIP Conference Proceedings*; Goyal, P., Giffin, A., Knuth, K.H., Vrscay, E., Eds.; American Institute of Physics: College Park, MD, USA, 2012; Volume 1443, pp. 230–237. [CrossRef]
14. Lary, D.J.; Faruque, F.S.; Malakar, N.; Moore, A.; Roscoe, B.; Adams, Z.L.; Eggelston, Y. Estimating the global abundance of ground level presence of particulate matter (PM 2.5). *Geospat. Health* **2014**, *8*, 611–630. [CrossRef]
15. Lary, D.; Lary, T.; Sattler, B. Using Machine Learning to Estimate Global PM2. 5 for Environmental Health Studies. *Environ. Health Insights* **2015**, *9*, 41.
16. Lary, D.J.; Alavi, A.H.; Gandomi, A.H.; Walker, A.L. Machine learning in geosciences and remote sensing. *Geosci. Front.* **2016**, *7*, 3–10. [CrossRef]
17. Kneen, M.A.; Lary, D.J.; Harrison, W.A.; Annegarn, H.J.; Brikowski, T.H. Interpretation of satellite retrievals of PM 2.5 over the Southern African Interior. *Atmos. Environ.* **2016**, *128*, 53–64. [CrossRef]
18. Liu, X.; Wu, D.; Zewdie, G.K.; Wijerante, L.; Timms, C.I.; Riley, A.; Levetin, E.; Lary, D.J. Using machine learning to estimate atmospheric Ambrosia pollen concentrations in Tulsa, OK. *Environ. Health Insights* **2017**, *11*, 1–10. [CrossRef]
19. Nathan, B.J.; Lary, D.J. Combining Domain Filling with a Self-Organizing Map to Analyze Multi-Species Hydrocarbon Signatures on a Regional Scale. *Environ. Model. Assess.* **2019**, *191*, 1–17. [CrossRef]
20. Lary, D.J.; Zewdie, G.K.; Liu, X.; Wu, D.; Levetin, E.; Allee, R.J.; Malakar, N.; Walker, A.; Mussa, H.; Mannino, A.; et al. Machine Learning Applications for Earth Observation. In *Earth Observation Open Science and Innovation*; ISSI Scientific Report Series; Springer: Berlin/Heidelberg, Germany, 2018; Volume 15, pp. 165–218.
21. Lary, M.A.; Allsop, L.; Lary, D.J. Using Machine Learning to Examine the Relationship between Asthma and Absenteeism. *Environ. Model. Assess.* **2019**, *191*, 1–9. [CrossRef]
22. Zewdie, G.K.; Lary, D.J.; Levetin, E.; Garuma, G.F. Applying deep neural networks and ensemble machine learning methods to forecast airborne ambrosia pollen. *Int. J. Environ. Res. Public Health* **2019**, *16*, 1992. [CrossRef]
23. Zewdie, G.K.; Lary, D.J.; Liu, X.; Wu, D.; Levetin, E. Estimating the daily pollen concentration in the atmosphere using machine learning and NEXRAD weather radar data. *Environ. Monit. Assess.* **2019**, *191*, 418. [CrossRef] [PubMed]
24. Wijeratne, L.O.; Kiv, D.R.; Aker, A.R.; Talebi, S.; Lary, D.J. Using Machine Learning for the Calibration of Airborne Particulate Sensors. *Sensors* **2020**, *20*, 99. [CrossRef] [PubMed]
25. Fingas, M.F.; Brown, C.E. Review of oil spill remote sensing, The Second International Symposium on Oil Spills. *Spill Sci. Technol. Bull.* **1997**, *4*, 199–208. [CrossRef]
26. Fingas, M. *Oil Spill Science and Technology*; Gulf Professional Publishing: Houston, TX, USA, 2010.

27. Liu, Y.; MacFadyen, A.; Ji, Z.; Weisberg, R. *Monitoring and Modeling the Deepwater Horizon Oil Spill: A Record Breaking Enterprise*; Geophysical Monograph Series; Wiley: Hoboken, NJ, USA, 2013.
28. Cornwall, W. Deepwater Horizon: After the oil. *Science* **2015**, *348*, 22–29. [CrossRef] [PubMed]
29. Breiman, L. Random Forests. *Mach. Learn.* **2004**, *45*, 5–32. [CrossRef]
30. Belgiu, M.; Dragut, L. Random forest in remote sensing: A review of applications and future directions. *Isprs J. Photogramm. Remote. Sens.* **2016**, *114*, 24–31. [CrossRef]
31. Probst, P.; Wright, M.N.; Boulesteix, A. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1301. [CrossRef]
32. Nocedal, J.; Wright, S. *Numerical Optimization*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006.
33. Aurin, D.; Maninno, A.; Lary, D.J. Remote Sensing of CDOM, CDOM Spectral Slope, and Dissolved Organic Carbon in the Global Ocean. *Appl. Sci.* **2018**, *8*, 2434. [CrossRef]

Advancement in Airborne Particulate Estimation Using Machine Learning

Lakitha Omal Harindha Wijerante, Gebreab K. Zewdie, Daniel Kiv, Adam Aker, David J. Lary, Shawhin Talebi, Xiaohe Yu, and Estelle Levetin

Introduction

The air we breathe is vital and largely invisible (except when the pollution levels are very high). Every single minute, an average human being breathes around 10 liters of air. However, we often do not think about the composition of the air that we breathe and the impact it may be having on our health. Often, the air we breathe contains pollutant particles. Although it is apparent that air pollution results in increased hospital visits, missed school days, as well as missed work days (due to respiratory diseases), it is harder to localize exactly where unhealthy air resides. The World Health Organization (WHO) reports that nine out of ten people worldwide breathe polluted air which results in an estimated 7 million deaths per year (Nada Osseiran 2018).

Air Pollution Episodes in History

Historically, we have seen that air pollution episodes can result in significant loss of life. A few of example episodes include:

- Great smog of London (1952): In December of 1952, a severe smog covered many parts of the British Isles (Wilkins 1954). The episode lasted 5 days (December 5–December 9, 1952), and more than 4000 deaths occurred before the end of the year. Within the next 10 weeks, a further 8000 people lost their lives (Black 2003). The primary cause of the episode was extensive burning of high-sulfur coal (Polivka 2018). Following the incident, the British Parliament passed the Clean Air Act of 1956 which restricted burning of coal in urban areas.
- New York City smog (1966): On the Thanksgiving weekend of 1966, smog containing damaging levels of toxic pollution (comprised of carbon monoxide and sulfur dioxide) covered New York City. Carlson (2009) reports that Thanksgiving weekend of 1966 was the smoggiest day in the city's history. Although regional leaders announced a first-stage alert, it is believed that more than 200 people lost their lives due to the air pollution episode. In fact, Glasser et al. (1967) estimate that 24 excess deaths per day occurred in New York City during the air pollution episode (November 23–November 29, 1966). In the wake of such environmental pollution events, as a means of limiting and eradicating environmental pollution, the US EPA (United States Environmental Protection Agency) was established on December of 1970.
- Eastern China smog (2013): On December 7, 2013, a hazardous smog stretched a distance of about 2 km within China (Levy 2014). The episode lasted for 8 days between the 2nd and 9th of December 2013. Huang et al. (2016) state that within the duration of the smog episode, the average $PM_{2.5}$ was 212 $\mu g/m^3$, which was three times higher than the usual $PM_{2.5}$ concentration (76 $\mu g/m^3$) within the same area. In China, coal still remains to be the main energy source, and it's regarded to be the primary cause of fine PM pollution in China. The Chinese cities of Baoding, Shijiazhuang, and Handan reported more than 30,000 deaths in 2013 per city, which can be linked to pollution (Solomon 2016).
- Great smog of New Delhi (2016): WHO announced New Delhi as the most polluted city in the world in 2014 (Saravanan et al. 2017). On November of 2017, air pollution levels in New Delhi went up to 999 on the AQI

L. O. H. Wijerante · G. K. Zewdie · D. Kiv · A. Aker · D. J. Lary (✉)
S. Talebi · X. Yu · E. Levetin
Hanson Center for Space Sciences, The University of Texas at Dallas, Richardson, TX, USA
e-mail: lhw150030@utdallas.edu; david.lary@utdallas.edu

F. S. Faruque (ed.), *Geospatial Technology for Human Well-Being and Health*,
https://doi.org/10.1007/978-3-030-71377-5_13

(Air Quality Index) scale. This is an air pollution level equivalent to smoking 50 cigarettes per day (Basu 2019). The cities' visibility level reduced to more than 50 m during the episode (Terry et al. 2018). It is assumed that the major causes of air pollution in New Delhi are burning coal, petrol, diesel, gas, biomass, and waste, along with industries, power plants, and firecrackers (Saravanan et al. 2017).

Such episodes remain a constant reminder of the devastation that air pollution can cause. The first step in fighting air pollution is to quantify the problem.

Making the Invisible, Visible

Conventional air quality management systems generally rely on a small number of regulatory-grade sensors across an urban area, for example, across the Dallas-Fort Worth Metroplex with a population of over seven million, there are just three airborne particulate sensors. Due to the substantial cost of these regulatory-grade sensing systems, they fail to provide adequate spatial and temporal resolution for characterizing air quality on a neighborhood scale. As such, these sensor systems do not adequately inform us about the situation within our neighborhoods, where people live, work, and play. Recent studies have demonstrated that air quality varies on very fine spatial and temporal scales (Harrison 2015; Harrison et al. 2015). As such, it is apparent that the world needs air quality sensing systems at the neighborhood scale (i.e., with a spatial resolution of less than a km). For example, a study that made near daily fine scale measurements at least every meter over a 100 km^2 in north Texas (Harrison 2015; Harrison et al. 2015) used variograms to characterize the spatial scale of airborne particulates. These studies revealed that the spatial scales over the study period depended on the synoptic situation and varied between 0.5 and 7.5 km.

The initiation of air quality sensing system requires an understanding of what parameters to measure, where the sensors are to be located, and the budgetary constraints that the system is to be bounded by. Such an understanding can be gained via the knowledge of the mix of pollutants that may be residing within the study and the general mindset of the people at stake. Another key aspect in getting started in a project is to look into available technologies that can abide by the requirements defined.

Airborne Particulates

Airborne atmospheric aerosols are an assortment of solid or liquid particles suspended in air (Boucher 2015). Aerosols, also referred to as particulate matter (PM), are associated with a suite of issues relevant to the global environment (Charlson et al. 1992; Ramanathan et al. 2001; Dubovik et al. 2002; Guenther et al. 2006; Hallquist et al. 2009; Kanakidou et al. 2005; Allen et al. 2014), atmospheric photolysis, and a range of adverse health effects (Dockery et al. 1993a; Oberdörster et al. 2005; Pope III et al. 2002; Pope et al. 2006; Cheng and Liu 2009; Chin 2009; Lim et al. 2012). Atmospheric aerosols are usually formed either by direct emission from a specific source (e.g., combustion) or from gaseous precursors (Stocker 2014). Although individual aerosols are typically invisible to the naked eye, due to their small size, their presence in the atmosphere in substantial quantities means that their presence is usually visible, e.g., as fog, mist, haze, smoke, dust plumes, etc. (Seinfeld 1986). Airborne aerosols vary in size, composition, and origin as well as in spatial and temporal distributions (Chin 2009; Pöschl 2005). As a result, the study of atmospheric aerosols has numerous challenges. The following aerosol classifications provide some useful insights.

Aerosol Classification

Characterization of atmospheric aerosols can be based on their origin, concentration, size, chemical composition, phase, and morphology (Seinfeld 1986). However, one of the main forms of aerosol classification is via their sources.

Source-Based Classification

The formation of atmospheric aerosols can be complex. As a result, the determination of global aerosol sources is approximate (Kondratyev et al. 2006). Three main terrestrial sources are typically quoted (Kokhanovsky 2008).

- Cosmic aerosols: Particles migrating through space are usually considered cosmic particles (Carslaw et al. 2002; Kirkby et al. 2011).
- Primary aerosols: Particles directly emanating from the earth's surface are usually termed primary aerosols (Holben et al. 2001; Streets et al. 2003; Bond et al. 2004; Kanakidou et al. 2005), for example, aerosols formed due to the agitation of oceanic or terrestrial surfaces by wind.
- Secondary aerosols: Secondary particles occur from condensation of gaseous species (aerosol precursors) (Atkinson 2000; Kanakidou et al. 2005; Hallquist et al. 2009; Jimenez et al. 2009). These may endure one or many chemical transformations prior their formation.

Primary and secondary aerosols are further subdivided depending on their origin into natural and anthropogenic (man-made) aerosols (Schauer et al. 1996; Andreae and Crutzen 1997; Yunker et al. 2002; Pöschl 2005; Kondratyev et al. 2006; Boucher 2015; Colbeck and Lazaridis 2010).

Most emissions from the oceans, vegetation, forest fires, and volcanoes are considered natural in origin. Anthropogenic sources are dominated by the emissions from the combustion of fossil and biofuels.

Shape-Based Classification

Colbeck (2014) describes the three main distinctions based on the shapes of atmospheric aerosols.

- Isometric particulates: The three dimensions of isometric particles are defined to be similar. Spherical particulates belong to this category (Wachs 2009). This study is done under the assumption of sphericity (isometric particulates).
- Platelets: Platelets have two longer dimensions compared to the third one. Disk-like particles fall under this classification.
- Fibers: Fibers are particulates with two smaller dimensions and one longer dimension. Asbestos is one well-known fiber.

Classification Based on Chemical Composition

Ambient PM is usually comprised of a mixture of one or more of the following chemical compounds: geological material (oxides of aluminum, silicon, calcium, titanium, iron, and other metal oxides), sulfates, nitrates, ammonium, sodium chloride, organic carbon, elementary carbon, and liquid water (Chow et al. 1998). A more generalized classification is derived from considering the chemical purity of PM by distinguishing between internal and external mixtures.

- External mixture: In an external mixture, individual particles within are chemically pure.
- Internal mixture: In an internal mixture, individual particulates are a mix of chemical species. A perfect internal mixture is said to have the same mix of chemical species for all particulates.

Typical atmospheric particulates would be in the middle ground between perfect internal and external mixtures (Boucher 2015). The optical properties of atmospheric aerosols, and in turn the radiative forcing due to atmospheric aerosols, are partly determined by the state of mixing (externally or internally) of the chemical species involved (Lesins et al. 2002).

Spatial Classification

Aerosols are also categorized with respect to their localized regions. The classification gives rise to these categories: urban aerosols, marine aerosols, rural continental aerosols, free troposphere aerosols, stratospheric aerosols, polar aerosols, and desert aerosols. In some cases, a geospatial classification might be inexact due to the possibility of long-range aerosol transportation. However, the regional aerosol classification is useful when local effects eclipse the more generic effects of aerosols (Boucher 2015).

Size-Based Classification

Aerosol size distribution and chemical composition play a role in their atmospheric transportation (Colbeck and Lazaridis 2010). Most atmospheric particles are not spherical. However, in atmospheric sciences, particles with equivalent settling velocities are considered to be of equal size irrespective of their actual size or composition. The microscopic properties of aerosols differ significantly depending on the type of aerosol. Nevertheless, generic models are defined to describe the main microscopic properties of a given aerosol with its appropriately assumed diameter (Kokhanovsky 2008). The two most generic definitions of such assumed diameters are as follows:

- Aerodynamic diameter: The diameter of a unit density sphere which has similar aerodynamic properties as the particle considered.
- Stokes diameter: The diameter of a sphere which has similar density as the particle considered.

These definitions are introduced to avoid ambiguities of size measurements that may occur due to using different types of instrumentation (Colbeck 2014). This study uses the aerodynamic diameter for size-based distinctions. There are two distinct means of aerosol classification with respect to size:

- Modal distributions: The size-based classification of aerosols is mainly devised on five modes (Boucher 2015; Alfarra 2004; Stier et al. 2005; Sỳkorová et al. 2016):
  1. The nucleation mode or ultrafine mode with a diameter of less than 0.01 μm.
  2. The Aitken mode with a diameter in the range 0.01 μm $-$ 0.1 μm.
  3. The accumulation mode with a diameter in the range 0.1 μm $-$ 1 μm.
  4. The coarse mode with a diameter in the range 1 μm $-$ 10 μm.
  5. The super-coarse mode with a diameter of greater than 10 μm.

  Each of these modes corresponds to the relative maximums of number, surface, and volume distributions of atmospheric aerosols.
- Variables related to human exposure: The term "fine" (or ultrafine) particulates usually refers to particulates less than 1 μm in aerodynamic diameter (PM_1) and particulates less than 2.5 μm in aerodynamic diameter ($PM_{2.5}$). For air pollution control, particulates up to 10 μm in diameter (PM_{10}) are also considered (Pöschl 2005). Cur-

rently, the US EPA (United States Environmental Protection Agency) regulates $PM_{2.5}$ and PM_{10} due to the human health effects associated with $PM_{2.5}$ and PM_{10} (US EPA 2004). Some air quality monitors also measure the total suspended particle (TSP) size fraction which includes particulates up to 40 μm (Chow et al. 1998). Another division of occupational health-based size-selective sampling is defined by assessing the subset of particles that can reach a selective region of the respiratory system. On this basis, three main fractions are defined: inhalable, thoracic, and respirable (Hinds 2012). The current study focuses on measurements of the six variables PM_1, $PM_{2.5}$, PM_{10}, respirable (alveolic), thoracic, and inhalable size fractions.

Health Context

The effects on human health due to air pollution may be the most controversial (Seinfeld 1986). Nevertheless, it is by far the most important. Studies have shown that exposure of excess particulate matter has alarming negative health effects (Mannucci 2017). The smallest size ranges of (less than 2.5 μm) PM is capable of penetrating through to the lungs or even to one's bloodstream. As such, the highest mortality is associated with $PM_{2.5}$ (Chen et al. 2011). HEI (2017) reports that more than 90% of the world's population lived with unhealthy air in 2015. The American Thoracic Society (ATS) has a slightly higher guideline of 11 $\mu g/m^3$ annual mean concentrations as compared with the WHO's 10 $\mu g/m^3$ for $PM_{2.5}$. However, it is reported that 14% of countries with valid design values for atmospheric pollution exceed the said recommendation by the ATS (Cromar et al. 2016). The results of the Aphekom project conducted in 25 European cities reveal that complying with the WHOs PM guidelines for $PM_{2.5}$ would increase life expectancy by 22 months while also giving financial savings of €31 billion annually (Pascal et al. 2013). The health hazard created by excess airborne PM also creates critical expenditures within developing countries. The estimated economic cost due to $PM_{2.5}$ pollution for the city of Delhi was estimated to be $6394.74 million in 2015, up from $2714.10 million in 2005 (Maji et al. 2017). Due to these reasons, considerable amounts on research are done on the health hazards caused by PM. Table 1 provides an overview of research done on specific health concerns with respect to PM_{10}, $PM_{2.5}$, and UFPs (ultrafine particles).

Aerosol concentration, size, structure, and chemical composition are key factors in driving the health outcomes caused. However, these parameters are highly irregular in temporal and spatial schemes (Pöschl 2005). As such, even though the effect of PM exposure can be substantial, predicting a link between PM and human health can be challenging. Most studies rely on obtaining the level of morbidity and mortality for a given disease which can be attributed to the exposure to PM. Some studies also employ questionnaires in collecting health-related data.

Long-term exposure to $PM_{2.5}$ increases the risk of total and cardiovascular disease (CVD) mortality. The study by Thurston et al. (2016) concludes that $PM_{2.5}$ exposure has a substantial association with both total mortality and CVD mortality, with CVD having the highest hazard ratio of 1.10 for the study set of participants between 50 and 71 years. Pope et al. (2004) state that a 10 $\mu g/m^3$ increase in fine PM results in an 8%–18% increase in the mortality risk. A study conducted in six cities across the United States with a total of 8111 participating adults found that fine particulate air pollution was linked with excess mortality (Dockery et al. 1993b).

The ATS report (Cromar et al. 2016) for 2011–2013 found that 26% out of 21,400 excess morbidities and 26% out of 9320 excess deaths were associated with elevated $PM_{2.5}$ in the United States per year. A European study (Boldo et al. 2006) estimated that a reduction of the $PM_{2.5}$ abundance by 15 $\mu g/m^3$ of $PM_{2.5}$ would prevent 16,926 deaths annually within a subset of 23 European cities and that such a reduction would likely increase the life expectancy between 1 month to more than 2 years. The study included major cities, London, Paris, Athens, Barcelona, Madrid, and Valencia. In the study, excess exposure to $PM_{2.5}$ was viewed as a modifiable factor which causes cardiovascular morbidity and mortality. Maji et al. (2017) found that the mortality in Mumbai and Delhi during 2015 was associated with PM_{10} and lead to 32,014 and 48,651 deaths, respectively.

Cerebrovascular accidents are a prominent cause of morbidity throughout the world. It was estimated that an increase of 10 $\mu g/m^3$ of $PM_{2.5}$ accounts for 1.29% (95% CI 0.552%–2.03%) increase in the risk of emergency hospital admissions (Santibañez et al. 2013). Sulfate aerosols are known to cause respiratory throat and fever symptoms (Onishi et al. 2018).

In some cases, the maternal exposure to excess particulate matter has resulted in lower birth weights (LBW). A multicountry evaluation of LBW reveals that a 10 $\mu g/m^3$ increase in PM_{10} (odds ratio (OR) = 1.03; 95% confidence interval (CI), 1.01–1.05) and $PM_{2.5}$ (OR = 1.10; 95% CI 1.03–1.18) exposure during the entire pregnancy is positively correlated with LBWs (Dadvand et al. 2013).

Excess PM exposure can also be behind excess stress among individuals. Evidence has been found that mitochondrially encoded TRNA phenylalanine (MT-TF) and mitochondrially encoded 12S RNA (MT-RNR1) is linked with metal-rich PM_1 (Byun et al. 2013). Both mitochondrial MT-TF and MT-RNR1 DNA methylation are sources of oxidative stress which responds to foreign environments. Short-term exposure to $PM_{2.5}$ also prompts a mechanism involving pulmonary oxidative stress which in turn induces vascular insulin resistance and inflammation (Haberzettl et al. 2016).

Table 1 Health Concerns due to PM 10, PM 2.5, and ultrafine particles (UFPs). Table adapted from Ruckerl et al. (2006)

Health outcomes	Short-term studies			Long-term studies		
	PM_{10}	$PM_{2.5}$	UFP	PM_{10}	$PM_{2.5}$	UFP
Mortality						
All causes	xxx	xxx	x	xx	xx	x
Cardiovascular	xxx	xxx	x	xx	xx	x
Pulmonary	xxx	xxx	x	xx	xx	x
Pulmonary effects						
Lung function, e.g., PEF	xxx	xxx	xx	xxx	xxx	
Lung function growth				xxx	xxx	x
Asthma and COPD exacerbation						
Acute respiratory symptoms		xx	x	xxx	xxx	
Medication use			x			
Hospital admission	xx	xxx		x	x	
Lung cancer						
Cohort				xx	xx	x
Hospital admission				xx	xx	x
Cardiovascular effects						
Autonomic nervous system	xxx	xxx		x	x	
ECG-related endpoints						
Autonomic nervous system	xxx	xxx	xx			
Myocardial substrate and vulnerability		xx	x			
Vascular function						
Blood pressure	xx	xxx	x			
Endothelial function	x	xx	x			
Blood markers						
Pro-inflammatory mediators	xx	xx	xx			
Coagulation blood markers	xx	xx	xx			
Diabetes	x	xx	x			
Endothelial function	x	x	xx			
Reproduction						
Premature birth	x	x				
Birth weight	xx	x				
IUR/SGA	x	x				
Fetal growth						
Premature birth	x					
Infant mortality	xx	x				
Sperm quality	x	x				
Neurotoxic effects						
Central nervous system		x	xx			

Notes: X, few studies (6 or less); XX, many studies (7–10); XXX, large number of studies (>10).
Abbreviations: UFP, ultrafine particle; PEF, peak expiratory flow; COPD, chronic obstructive pulmonary disease; IUG, intrauterine growth restriction; SGA, small for gestational age

Environmental pollution is a potential cause of lung cancer. Tandem repeats are DNA sequences which lie adjacent to each other in the same orientation (direct tandem repeats) or in the opposite direction to each other. These DNA sequences are generally hypomethylated in cancer patients. A case study done on two contrasting groups on air pollution exposure of truck drivers and office workers reveals that PM is linked with hypomethylation of some tandem repeats (SATα, NBL2) (Guo et al. 2014).

The most likely candidates to be affected by unhealthy air are the elderly and infants. Pun et al. (2017) concluded that $PM_{2.5}$ is linked to both depressive and anxiety symptoms within older adults with the strongest association to individuals with lower socioeconomic measures. Shy et al. (1973) confirm that school children between the age of 9 and

13 exposed to elevated air pollution experience ventilatory problems. A study conducted with a collection of 40 fifth grade school children revealed that the "soot" fraction of $PM_{2.5}$ is strongly linked with pollution-related asthma attacks affecting children residing beside roadways (Spira-Cohen et al. 2011).

Using a business-as-usual emission scenario model, (Lelieveld et al. 2015) estimate that premature mortality due to outdoor air pollution could double by 2050. As such, it is of utmost importance to conduct in-depth research on PM and other air pollutant sources in order to enforce proper air pollution policies (Kelly and Fussell 2016).

Difficulty in Estimating Airborne Particulates

Conventional regulatory-grade instrumentation is accurate, but expensive. This makes it challenging to provide neighborhood-scale measurements due to the substantial costs involved. So in this study, we present two different case studies where we use machine learning to utilize different sensor types. First, we use low-cost optical particle counters that can be deployed at scale across neighborhoods. Second, we use remotely sensed observations made using weather RADARs.

What Is Machine Learning?

Machine learning has already proved useful in a wide variety of applications in science, business, healthcare, and engineering. Machine learning allows us to *learn by example* and to *give our data a voice*. It is particularly useful for those applications for which we do *not* have a complete theory, yet which are of significance. Machine learning is an automated implementation of the scientific method (Domingos 2015), following the same process of generating, testing, and discarding or refining hypotheses. While a scientist or engineer may spend his entire career coming up with and testing a few hundred hypotheses, a machine learning system can do the same in a fraction of a second. Machine learning provides an objective set of tools for automating discovery. It is therefore not surprising that machine learning is currently revolutionizing many areas of science, technology, business, and medicine (Lary et al. 2016, 2018).

Machine learning is now being routinely used to work with large volumes of data in a variety of formats such as image, video, sensor, health records, etc. Machine learning can be used in understanding this data and creating predictive and classification tools. When machine learning is used for regression, empirical models are built to predict continuous data, facilitating the prediction of future data points, e.g., algorithmic trading and electricity load forecasting. When machine learning is used for classification, empirical models are built to classify the data into different categories, aiding in the more accurate analysis and visualization of the data. Applications of classification include facial recognition, credit scoring, and cancer detection. When machine learning is used for clustering, or unsupervised classification, it aids in finding the natural groupings and patterns in data. Applications of clustering include medical imaging, object recognition, and pattern mining. Object recognition is a process for identifying a specific object in a digital image or video. Object recognition algorithms rely on matching, learning, or pattern recognition algorithms using appearance-based or feature-based techniques. These technologies are being used for applications such as driver-less cars, automated skin cancer detection, etc.

Machine learning is an automated approach to building empirical models from the data *alone*. A key advantage of this is that we make *no* a priori assumptions about the data, its functional form, or probability distributions. It is an empirical approach. However, it also means that for machine learning to provide the best performance, we do need a *comprehensive representative set of examples*, which spans as much of the parameter space as possible. This comprehensive set of examples is referred to as the *training data*.

So, for a successful application of machine learning, we have *two* key ingredients, both of which are essential, a machine learning algorithm and a comprehensive training dataset. Then, once the training has been performed, we should test its efficacy using an independent validation dataset to see how well it performs when presented with data that the algorithm has *not* previously seen, i.e., test its *generalization*. This can be, for example, a randomly selected subset of the training data that was held back and then utilized for independent validation.

It should be noted that with a given machine learning algorithm, the performance can go from poor to outstanding with the provision of a progressively more complete training dataset. Machine learning really is learning by example, so it is critical to provide as complete a training dataset as possible. At times, this can be a labor-intensive endeavor.

A key part of machine learning studies is an independent validation to objectively test the "generalization" of the empirical models. This is often done by randomly splitting the available data into two portions. One portion, the training dataset, is used to train the empirical machine learning model. The other portion, the independent validation dataset, is used to objectively test the empirical model by using data not seen in the training process.

We have used machine learning in many previous studies (Brown et al. 2008; Lary et al. 2009a; Lary and Aulov 2008; Lary et al. 2004; Malakar et al. 2013; Lary 2010; Malakar et al. 2012a; Lary 2013, 2007; Albayrak et al. 2011; Lary et al. 2003; Malakar et al. 2012b; Lary 2014; Lary et al.

2015b; Kneen et al. 2016; Lary et al. 2010; Medvedev et al. 2016; Lary et al. 2016; O et al. 2017; Wu et al. 2017; Nathan and Lary 2019; Lary et al. 2019, 2018; Wu et al. 2019; Alavi et al. 2016; Ahmad et al. 2016; Zewdie and Lary 2018; Malakar et al. 2018; Zewdie et al. 2019a,b; Chang et al. 2019; Choi et al. 2019). In this study, we have used machine learning for multivariate nonlinear non-parametric regression. Some of the commonly used regression algorithms include neural networks (McCulloch and Pitts 1943; Haykin 2001, 2007, 1994, 1999; Demuth et al. 2014; Bishop 1995), support vector machines (Vapnik 1982, 1995; Cortes and Vapnik 1995; Vapnik 2000, 2006), decision trees (Safavian and Landgrebe 1991), and ensembles of trees such as random forests (Ho 1998; Breiman 1984, 2001). Previously, we have used a similar approach to cross-calibrate satellite instruments (Lary and Aulov 2008; Brown et al. 2008; Lary et al. 2009a, 2016, 2018). Recently, other studies have also used machine learning to calibrate low-cost sensors (Li et al. 2014; Dong et al. 2015).

Case study: Using Machine Learning for the Calibration of Airborne Particulate Sensors

Low-cost sensors that can also be accurately calibrated are of particular value. For the last two decades, we have pioneered the use of machine learning to cross-calibrate sensors of all kinds. This was initially done for very expensive orbital instruments onboard satellites (awarded an IEEE paper prize and specially commended by the NASA MODIS team) (Lary et al. 2009a). We are now using this approach operationally for low-cost sensors distributed at scale across dense urban environments as part of our smart city sentinels. The approach can be used for very diverse sensors, but as a useful illustrative example that has operational utility, we describe here a use case for accurately calibrated low-cost sensors measuring the abundance and size distribution of airborne particulates, with the implicit understanding that many other sensor types could easily be substituted. These sensors can be readily deployed at scale at fixed locations, mobile on various robotic platforms (walking, flying, etc.) or vehicles, carried, or deployed autonomously as a mesh network, either by operatives or by robots (walking, flying, etc.).

Building-in calibration will enable consistent data to be retrieved from all the low-cost sensors. Otherwise, the data will always be under some suspicion as the inter-sensor variability among low-cost nodes can be substantial. While much effort has been recently placed on providing the connectivity of large disbursed low-cost networks, little to no effort has been spent on the automated calibration, bias-detection, and uncertainty estimation necessary to make sure the information collected is sound. A case study of providing this critical calibration using machine learning is the focus of this paper.

Any sensor system benefits from calibration, but low-cost sensors are typically in particular need of calibration. The inter-sensor variability among low-cost nodes can be substantial. In addition to the pre-deployment calibration, once the sensors have been deployed, the paradigm we first developed for satellite validation of constructing probability distribution functions of each sensor's observation streams can be used to both monitor the real-time calibration of each sensor in the network by comparing its readings to those of its neighbors and also answer the question "how representative is an instantaneous reading of the conditions seen over some temporal and spatial window within which the sensor is placed?"

Using Probability Distribution Functions to Monitor Calibration and Representativeness in Real Time

It is useful to be able to answer the question "how representative is an instantaneous reading of the conditions seen over some temporal and spatial window within which the sensor is placed?" We can answer this question by considering a probability distribution function (PDF) of all the observations made by a sensor over some temporal and spatial window. The width of this probability distribution is termed the representativeness uncertainty for that temporal and spatial window. The PDFs of all observations made by each sensor are automatically compared in real time to the PDFs from the neighboring sensors within a neighborhood radius. These neighborhood sensors can include measurements from primary reference sensors that may be available. This approach is used to estimate the measurement uncertainty and inter-instrument bias for the last hour, day, etc. We continuously accumulate the PDF for each sensor over a variety of time scales and compare it to its nearest neighbors within a neighborhood radius. Any calibration drift in a sensor will be quickly identified as part of the fully automated real-time workflow where we will automatically be comparing each sensor's PDFs to its neighbor's PDFs and to the reference instruments PDFs. As each sensor is in a slightly different local environment, the sensor bias drift for each sensor will be different.

Characterizing the Temporal and Spatial Scales of Urban Air Pollution

This study focused on the calibration of low-cost sensors as part of a larger endeavor with the goal of characterizing the temporal and spatial scales of urban pollution. The temporal

and spatial scales of each atmospheric component are intimately connected. The resolution used in atmospheric chemistry modeling tools is often driven by the computational resources available. The spatial resolution of observational networks is often determined by the fiscal resources available. It is worth taking a step back and characterizing what the actual spatial scales are for each chemical component of urban atmospheric chemistry. Based on our street-level surveys providing data at less than a meter resolution, it is clear that the spatial scales are dependent on several factors such as the synoptic situation, the distribution of sources, the terrain, etc. In the larger study, we characterize the spatial scales of multi-species urban pollution by using a hierarchy of measurement capabilities that include (1) a zero emission electric survey vehicle with comprehensive gas, particulate, irradiance, and ionizing radiation sensing and (2) an ensemble of more than 100 street-level sensors making measurements every few seconds of a variety of gases, particulates, light levels, temperature, pressure, and humidity. Each sensor is accurately calibrated against a reference standard using machine learning. This paper documents an example of low-cost sensor calibration for airborne particulate observations.

Societal Relevance

What are the characteristic spatial scales of each chemical species, and how does this depend on issues such as the synoptic situation? These are basic questions that are helpful to quantify when considering atmospheric chemistry, when looking forward to the next generation of modeling tools and observing system (whether from space or ground-based networks), and when evaluating mitigation strategies, especially with regard to co-benefits for air pollution and greenhouse gas reduction and investigating the evolution of urban air composition in a warming climate. To be able to quantify these spatial and temporal scales, we need a comprehensive observing system; being able to use low-cost sensors is of great assistance in achieving this goal.

The Dallas Fort Worth (DFW) Metroplex (where our study was conducted) is the largest inland urban area in the United States and the nation's fourth largest metropolitan area. Nearly a third of Texans, more than seven million inhabitants, live in the DFW area. A population which is growing by a thousand people every day. DFW is an area with an interesting variety of specific pollution sources with unique signatures that can provide a useful testbed for generalizing a measurement strategy for dense urban environments. For more than two decades, the DFW area has been in continuous violation of the Clean Air Act. DFW will be one of only ten non-California metropolitan areas still in violation of the Clean Air Act in 2025 unless major changes take place. This has already had a detrimental health impact, e.g., even though the Texas average childhood asthma rate is 7%, and the national average is 9%, the DFW childhood asthma rate is 20–25%. Second only to the Northeast, DFW ranks second in the number of annual deaths due to smog. Further, a leading factor in poor learning outcomes in high schools is absenteeism, a leading cause of absenteeism is asthma, and key trigger for asthma is airborne pollution (Lary et al. 2019). Physical exertion in the presence of high pollution levels is more likely to lead to an asthma event. The sensors calibrated in this study are being provided to high schools and high school coaches so that simple practical decisions can be made to reduce adverse health outcomes, e.g., given the levels of pollen/pollution today, should physical education/practice be outside or inside?

The Datasets Used

All of the measurements were made at our own field calibration station in the ambient environment. The calibration of the low-cost AlphaSense OPC occurs prior to their deployment across the dense urban environment of DFW. In this study, we use machine learning to bring together two distinct types of data. First, we use accurate in situ observations made by a research-grade particulate spectrometer. Second, we use observations from inexpensive optical particle counters. The inexpensive sensors are particularly useful as they can be readily deployed at scale.

Research-Grade Optical Particle Counter

The particulate spectrometer is a laser-based optical particle counter (OPC). In this study we used a GRIMM Laser Aerosol Spectrometer and Dust Monitor Model 1.109. The sensor has the capability of measuring particulates of diameters between 0.25 μm and 32 μm distributed within 32 size channels. Such a wide range of diameter space is made possible due to intensity modulation of the laser source. Particulates pumped into the sensor are detected through scattering a laser beam of 655 nm into a light trap. The laser beam is aimed at particulates coming through a sensing chamber at a flow rate of 1.21 l/min. The device classifies particulates into specific size classes subject to its intensity (Broich et al. 2012). The optical arrangement of the sensor is staged such that a curved optical mirror placed at an average scattering angle of 90° collects and redirects the scattered light toward a photo sensor. The wide angle of the optical mirror (120°) is meant to increase the light intensity redirected toward the photo sensor within the Rayleigh scattering domain which decreases the minimum detectable particle size. Furthermore, it compensates for Mie scattering undulations caused by monochromatic illumination. The sensing period

of the GRIMM sensor was set to 6 s and for each time window provides three standardized mass fractions, namely, based on occupational health (repairable, thoracic, and alveolic) according to EN 481 as well as PM_1, $PM_{2.5}$, and PM_{10}.

Low-Cost Optical Particle Counters

There are several readily available optical particle counters (OPC) which are useful, but much less accurate compared to research grade sensors. In this study, we focus on using such sensors, together with machine learning, to get as close as possible to the accuracy of research-grade PM sensors. After the application of the machine learning calibration, these lower-cost sensors perform admirably. In order for low-cost sensors to provide an improved picture of PM levels, a careful calibration is required. The current study uses an Alpha Sense OPCN3 (http://www.alphasense.com/) together with a cheaper environmental sensor (Bosch BME280) as data collectors. The OPC-N3 is compact, 75 mm × 60 mm × 65 mm in size, and weighs under 105 g which uses similar technology to the conventional OPCs where particle size is determined via a calibration based on Mie scattering. Unlike most OPCs, the OPC-N3 doesn't include a pump and a replaceable particle filter in order to pump aerosol samples through a narrow inlet tube, hence avoiding the need for regular maintenance. Sufficient airflow through the sensor is made possible with a low-powered micro-fan producing a sample flow rate of 280 mL/min. The OPC-N3 is capable of onboard data logging as well as measuring particulates of diameters up to 40 μm. This enables the OPC-N3 to measure pollen and other biological particulates. The onboard data is saved within an SD card which can be accessed through a micro-USB cable connected to the OPC. Furthermore, the OPC-N3's lower sensing diameter is reduced to 0.35 μm as opposed to its predecessor's (OPC-N2) lower limit of 0.38 μm. The wider range of sensing is made possible via the OPC switching between high and low gain modes automatically. The OPC-N3 calculates its PM values using the method defined by the European Standard EN 481 (Alphasense 2018).

Caveat: Particulate Refractive Index

The observations made by optical particle counters are sensitive to the refractive index of the particulates and their light absorbing properties. The retrieved size distributions and the mass concentrations can be biased, depending on the nature of the particulates. The current study does not explore the accuracy implications of this. A future study is underway which includes direct measurements of black carbon that will allow us to begin to explore these aspects. The machine learning paradigm is readily extensible to include these aspects, even though not explicitly addressed in this study.

Machine learning is an ideal approach for the calibration of lower-cost optical particle counters.

Ensemble Machine Learning

Multiple approaches for nonlinear non-parametric machine learning were tried including neural networks, support vector regression, and ensembles of decision trees. The best performance was found using an ensemble of decision trees with hyperparameter optimization (Safavian and Landgrebe 1991; Ho 1998; Breiman 1984, 2001). Ensemble methods use multiple learners to obtain better predictive performance that could be obtained from any of the individual learners alone. A good example of an ensemble of learners is a random forest, which uses an ensemble of decision trees. In this study, the specific implementation used was that provided by the MathWorks in the fitrensemble function which is part of the MATLAB Statistics and Machine Learning Toolbox. Hyperparameter optimization was used so that the optimal choice was made for the following attributes: learning method (bagging or boosting), maximum number of learning cycles, learning rate, minimum leaf size, maximum number of splits, and the number of variables to sample. During hyperparameter optimization, we use an optimization approach (e.g., Bayesian optimization) to choose a set of optimal hyperparameters for our learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process.

In this study, there were 72 inputs to our multivariate nonlinear non-parametric machine learning regression; these include the particle counts for each of the 24 size bins measured by the OPC-N3; the OPC-N3 estimates of PM_1, $PM_{2.5}$, and PM_{10}; a suite of OPC performance variables including the reject ratio; and particularly important, the ambient atmospheric pressure, temperature, and humidity. The OPC-N3 sensor includes two photodiodes that record voltages which are eventually translated into particle count data. However, particles which are not entirely in the OPC-N3 laser beam, or are passing down the edge, are rejected, and this is recorded in the "reject ratio" parameter. This leads to better sizing of particles and hence plays an important role within the machine learning calibration.

Each of the six outputs we wished to estimate had its own empirical model. The performance of these six models in their independent validation is shown in Figs. 1 and 2. The outputs we estimated were the six variables measured by the reference instrument, the research-grade optical particle counter, namely, PM_1, $PM_{2.5}$, and PM_{10}, and the standardized occupational health respirable, thoracic, and alveolic mass fractions. The alveolic fraction is the mass fraction of in-

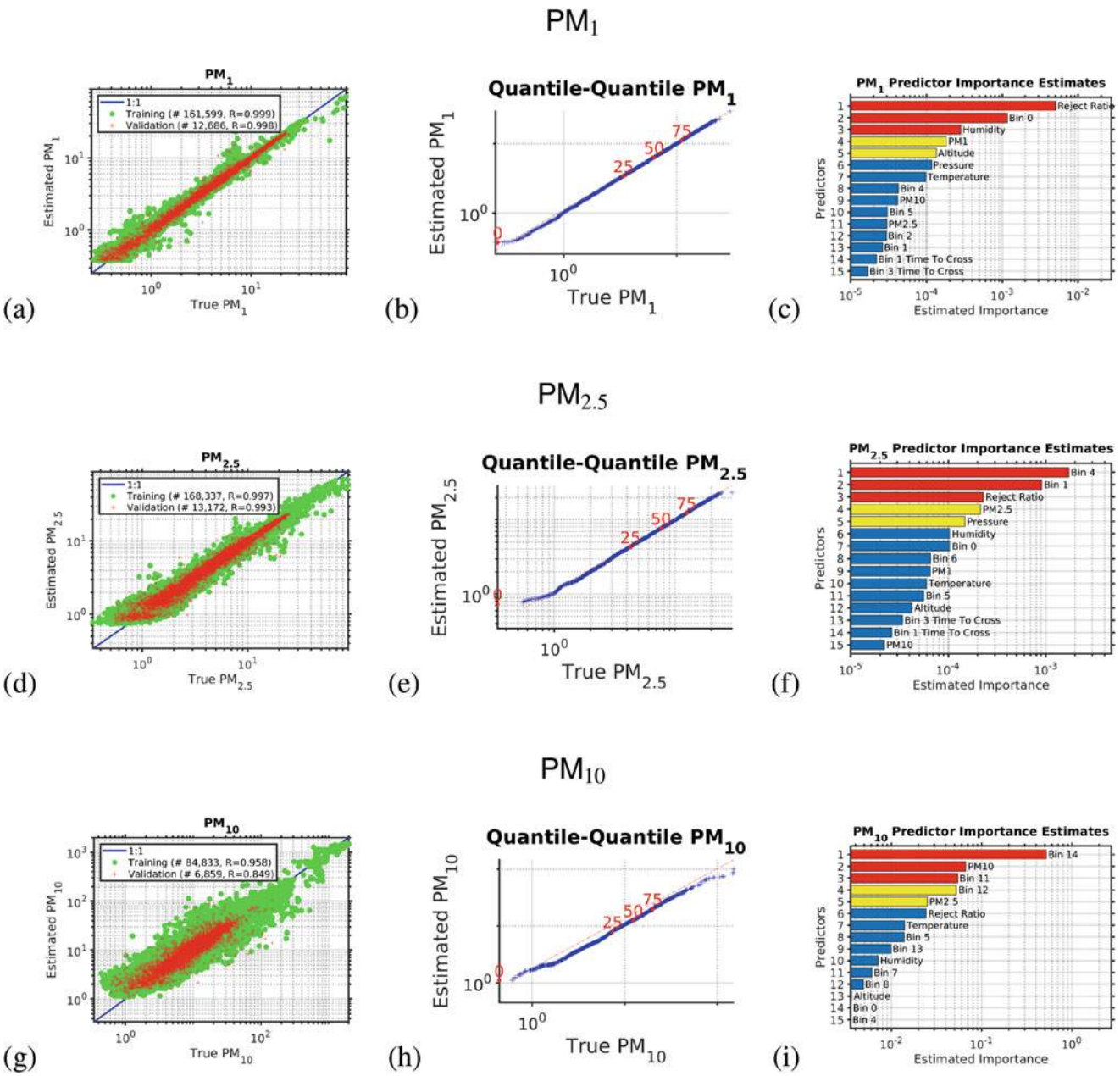


Fig. 1 This figure shows the results of the multivariate nonlinear non-parametric machine learning regression for PM_1 (panels **a**–**c**), $PM_{2.5}$ (panels **d**–**f**), and PM_{10} (panels **g**–**i**). The left-hand column of plots shows log-log axis scatter diagrams with the x-axis showing the PM abundance from the expensive reference instrument and the y-axis showing the PM abundance provided by calibrating the low-cost instrument using machine learning. The green circles are the training data, and the red pluses are the independent validation data. The blue line shows the ideal response. The middle column of plots shows the quantile-quantile plots for the machine learning validation data, with the x-axis showing the percentiles from the probability distribution function of the PM abundance from the expensive reference instrument and the y-axis showing the percentiles from the probability distribution function of the estimated PM abundance provided by calibrating the low-cost instrument using machine learning. The dotted red line shows the ideal response. The right-hand column of plots shows the relative importance of the input variables for calibrating the low-cost optical particle counters using machine learning.

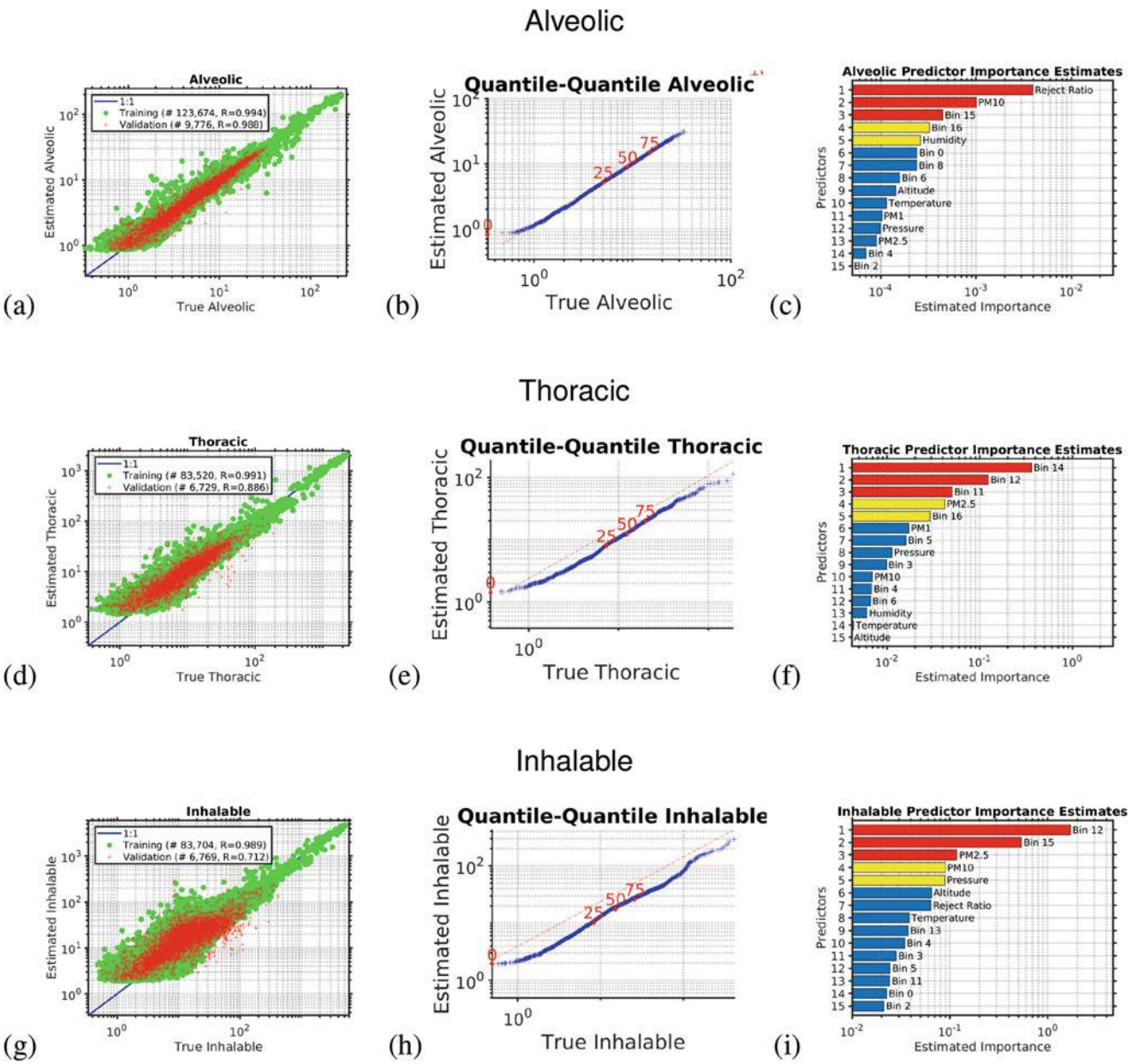


Fig. 2 This figure shows the results of the multivariate nonlinear non-parametric machine learning regression for the alveolic (panels **a**–**c**), thoracic (panels **d**–**f**), and inhalable size fractions (panels **g**–**i**). The left-hand column of plots shows log-log axis scatter diagrams with the x-axis showing the PM abundance from the expensive reference instrument and the y-axis showing the PM abundance provided by calibrating the low-cost instrument using machine learning. The green circles are the training data, and the red pluses are the independent validation dataset. The blue line shows the ideal response. The middle column of plots shows the quantile-quantile plots for the machine learning validation data, with the x-axis showing the percentiles from the probability distribution function of the PM abundance from the expensive reference instrument and the y-axis showing the percentiles from the probability distribution function of the estimated PM abundance provided by calibrating the low-cost instrument using machine learning. The dotted red line shows the ideal response. The right-hand column of plots shows the relative importance of the input variables for calibrating the low-cost optical particle counters using machine learning.

haled particles penetrating to the alveolar region (maximum deposition of particles with a size $\approx 2\,\mu m$). The thoracic fraction is the mass fraction of inhaled particles penetrating beyond the larynx ($<10\,\mu m$). The respirable fraction is the mass fraction of inhaled particles penetrating to the unciliated airways ($<4\,\mu m$). The inhalable fraction is the mass fraction of total airborne particles which is inhaled through the nose and mouth ($<20\,\mu m$). For each of these six parameters, we created an empirical multivariate nonlinear non-parametric machine learning regression model with hyperparameter optimization.

Calibrating the Low-Cost Optical Particle Counters Using Machine Learning

Figure 1 shows the results of the multivariate nonlinear non-parametric machine learning regression for PM_1 (panels a to c), $PM_{2.5}$ (panels d to f), and PM_{10} (panels g to i). The left-hand column of plots shows log-log axis scatter diagrams with the x-axis showing the PM abundance from the expensive reference instrument and the y-axis showing the PM abundance provided by calibrating the low-cost instrument using machine learning.

For the left-hand column of plots in Fig. 1 (the scatter diagrams), for a perfect calibration, the scatter plot would be a straight line with a slope of 1 and a y-axis intercept of 0; the blue line shows the ideal response. We can see that multivariate nonlinear non-parametric machine learning regression that we have used in this study employing an ensemble of decision trees with hyperparameter optimization has performed very well (panels a, d, and g). In each scatter diagram, the green circles are the data used to train the ensemble of decision trees, and the red pluses are the independent validation data used to test the generalization of the machine learning model.

We can see that the performance is best for the smaller particles that stay lofted in the air for a long period and do not rapidly sediment, so when comparing the scatter diagram correlation coefficients, r, for the independent validation test data (red points), we see that $r_{PM_1} > r_{PM_{2.5}} > r_{PM_{10}}$.

For the middle column of plots in Fig. 1 (the quantile-quantile plots), we are comparing the *shape* of the probability distribution (PDF) of all the PM abundance data collected by the expensive reference instrument to that of the the PM abundance provided by calibrating the low-cost instrument using machine learning. A $\log_{10}$ scale is used with a tick mark every decade. The dotted red line in each quantile-quantile plot shows the ideal response. The red numbers indicate the percentiles (0, 25, 50, 75, 100). If the quantile-quantile plot is a straight line, that means both PDFs have *exactly* the same shape as we are plotting the percentiles of one PDF against the percentiles of the other PDF. Usually we would like to see a straight line at least between the 25th and 75th percentiles; in this case, we have a straight line over the entire PDF, which demonstrates that the machine learning calibration has performed well.

The right-hand column of plots shows the relative importance of the input variables for calibrating the low-cost optical particle counters using machine learning. The relative importance metric is a measure of the error that results if that input variable is omitted. In the right-hand column of bar plots, we have sorted the importance metric into descending order, so the variable represented by the uppermost bar in each case was the most important variable for performing the calibration, the second bar is the second most important, etc. We note that along with the number of particles counted in each size bin, it is important to measure the temperature, pressure, and humidity to be able to accurately calibrate the low-cost OPC against the reference instrument. The data also suggests that the parameter "reject ratio" carries a higher deal of importance with respect to the calibration. OPC-N3 comprises two photodiodes which record voltages eventually translated into particle count data. However, particles which are not entirely in the beam or are passing down the edge are rejected and reflected on the parameter "reject ratio." This leads to better sizing of particles and hence plays a vital role within the ML calibration.

Another division of occupational health based size-selective sampling is defined by assessing the subset of particles that can reach a selective region of the respiratory system. On this basis three main fractions are defined: inhalable, thoracic, and respirable (Bickis 1998; Hinds 2012; Brown et al. 2013). Studies have shown that exposure of excess particulate matter has alarming negative health effects (Mannucci 2017). The smallest size ranges of particulate matter are capable of penetrating through to the lungs or even to one's bloodstream.

Figure 2 is similar to Fig. 1 and shows the results of the multivariate nonlinear non-parametric machine learning regression for the alveolic, thoracic, and inhalable size fractions. As would be expected, we see that the performance is best for the smaller particles that stay lofted in the air for a long period and do not rapidly sediment, so when comparing the scatter diagram correlation coefficients, r, for the independent validation test data (red points), we see that $r_{Alveolic} > r_{Thoracic} > r_{Inhalable}$.

Operational Use of the Calibration and Periodic Validation Updates

The calibration just described occurs pre-deployment of the sensors into the dense urban environment. Once these initial field calibration measurements are made over a period of several months, in the manner described above, the multi-

variate nonlinear non-parametric empirical machine learning model is applied in real time to the live stream of observations coming from each of our air quality sensors deployed across the dense urban environment of the Dallas Fort Worth Metroplex. These corrected measurements are then made publicly available as Open Data as well as depicted on a live map and dashboard.

Building-in continual calibration to a network of sensors will enable long-term, consistent, and reliable data. While much effort has been recently placed on the connectivity of large disbursed IoT networks, little to no effort has been spent on the automated calibration, bias detection, and uncertainty estimation necessary to make sure the information collected is sound. This is one of our primary goals. This is based on extensive previous work funded by NASA for satellite validation.

After deployment, a zero emission electric car carrying our reference is used, to routinely drive past all the deployed sensors to provide ongoing routine calibration and validation. An electric vehicle does not contribute any ambient emissions and so is an ideal mobile platform for our reference instruments.

For optimal performance, the implementation combines edge and cloud computing. Each sensor node takes a measurement at least every 10 s. The observations are continually time-stamped at the nodes and streamed to our cloud server, the central server aggregating all the data from the nodes and managing them. To prevent data loss, the sensor nodes store any values that have not been transmitted to the cloud server for reasons, including communication interruptions, in a persistent buffer. The local buffer is emptied to the cloud server at the next available opportunity.

Data from all sensors are archived and serve as an open dataset that can be publicly accessed. The observed probability distribution functions (PDFs) from each sensor are automatically compared in real time to the PDFs from the neighboring sensors within a neighborhood radius. These neighborhood sensors include measurements from the electric car/mobile validation sensors. This comparison is used to estimate the size-resolved measurement uncertainty and size-resolved inter-instrument bias for the last hour, day, week, month, and year. We continuously accumulate the PDF for each sensor over a variety of time scales (an hour, day, week, month, and year) and compare it to its nearest neighbors within a neighborhood radius.

Any calibration drift in a sensor will be quickly identified as part of a fully automated real-time workflow, where we will automatically be comparing each sensor's PDFs to its neighbor's PDFs and to the reference instruments PDFs. As each sensor is in a slightly different local environment, the sensor bias drift for each sensor will be different. We have previously shown that machine learning can be used to effectively correct these inter-sensor biases (Lary et al. 2009b). As a result, the overall distributed sensing system will not just be better characterized in terms of its uncertainty and bias but also provide improved measurement stability over time.

Case Study: Using Weather Radars and Machine Learning to Estimate Airborne Particulates

The application of radar for atmospheric meteorology started soon after the end of the Second World War. It was during the Second World War in the 1930s that radar technology was first used to locate and track war planes. The interference from scatterers such as rainfall prompted the notion that radar can also be applied to measure atmospheric precipitation. Subsequently, the construction of radar networks for meteorologic purposes commenced. The first radar network for meteorologic purposes was the Weather Surveillance Radar-1957 (WSR-57) in the United States. Currently, the WSR radar network has been upgraded to WSR-88D (Weather Surveillance Radar, 1988). WSR-88D has about 160 Doppler radars all over the United States. Technically WSR-88D radar is known as the Next-Generation Radar (NEXRAD). The following sections present the measurements of the NEXRAD radar and its application to identify aerosols.

Weather radars are mainly designed for determining and forecasting atmospheric phenomena such as precipitation, cloud coverage, wind direction and magnitude, and other associated meteorological events. In addition to these daily atmospheric conditions, radar can detect other objects and particles of small size such as dust, sand, insects, bird migrations, ground clutter, etc. The weather radar can also detect variations in the refractive index of the atmosphere caused by variations in the ambient temperature.

Atmospheric radars employed for meteorologic purposes transmit electromagnetic pulses of various frequencies. The frequency range used in the design of the radar determines the purpose and observation capability of the radar. For example, radars designed for observing the amount, type, and motion of precipitation have frequencies from 3–10 GHz (in terms of wavelength, 10–3 cm, respectively). Radars having this frequency range are very convenient for meteorological purposes. Radars having higher frequencies are useful to observe small-size droplets and particles. Small-size cloud particles, light snow, fog, and light rainfall are observed by high-frequency meteorological radars. At relatively low frequencies (in the range of less than 100–1000 MHz), the radar can detect fluctuations in the refractive index of the clear atmosphere. Low-frequency radars are best suited for profiling wind speed and direction.

The NEXRAD radar is in general operated in two different modes based on atmospheric weather conditions. These two

modes are the precipitation mode and the clear air mode. In the precipitation mode, the NEXRAD radar is operated at fast rotations at various elevations up to about 19.5°. In precipitation mode, a high emphasis is given for measurements at several elevations in order to see vertical storm profiles. In clear air mode, the radar is operated slowly, and it is sensitive to observe scattering from small objects such as pollen, other particulate matter, dust, smoke, insects, and birds (Gali 2010). The approximate time for a volume scan is 6 and 10 min. for precipitation and clear air modes, respectively.

Direct measurements of pollen and other particulates are rarely done using the NEXRAD radar. However, a few exceptional research projects have been reported showing observation of large aerosols using the NEXRAD weather radar (Madonna et al. 2010). Consequently radar scattering from large aerosols such as pollen is hard to identify. But NEXRAD measurements of Doppler velocity, direction, and speed of wind which are meteorological variables controlling the distribution and dispersal of pollen and large particulate matter. Other meteorological variables such as cloud coverage, precipitation, and rainfall are also pollen-controlling variables associated with the radar base reflectivity. For example, Eq. (1) shows the rainfall estimation techniques based on NEXRAD reflectivity.

$$Z = aR^b \tag{1}$$

where Z and R, respectively, represent reflectivity and rainfall and a and b are experimentally determined constants. a and b are determined experimentally comparing radar reflectivity and rain gauge measurements. The National Weather Service default value of a and b are 300 and 1.4, respectively.

The lack of a complete functional relationship between NEXRAD measurements and airborne particulates motivates us to seek other options. The machine learning approach of "learning" by example from large datasets is the perfect candidate for this problem. In machine learning, we estimate a variable based on a large number of input variables (data), and the method is becoming popular in a wide variety of fields.

In this study, the inputs to our multivariate nonlinear non-parametric machine learning regression were the remotely sensed parameters provided by the weather radar. The outputs we wished to estimate were the variables measured by the in situ optical particle counter.

Estimating Aerosol Size Distribution

Figure 3 shows the results of the multivariate nonlinear non-parametric machine learning regression as a function of particle size. The x-axis shows the particle size on a log scale. The y-axis shows the quality of fit using the correlation coefficient of the scatter diagram for each particle size fraction; a perfect fit would have a correlation coefficient of 1. The blue line shows the results for the training data. The red line shows the results for the independent validation data. We can see that the machine learning can effectively use the NEXRAD data for the particles with a size of less than 7 μm.

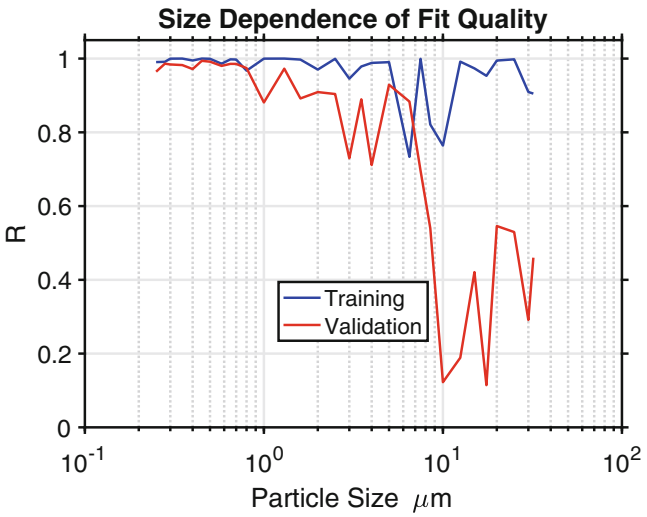


Fig. 3 This figure shows the results of the multivariate nonlinear non-parametric machine learning regression as a function of particle size. The x-axis shows the particle size on a log scale. The y-axis shows the quality of fit using the correlation coefficient of the scatter diagram for each particle size fraction; a perfect fit would have a correlation coefficient of 1. The blue line shows the results for the training data. The red line shows the results for the independent validation data. We can see that the machine learning can effectively use the NEXRAD data for the particles with a size of less than 7 μm

We can see a little more detail in Fig. 4 which shows the results of using an ensemble of regression trees for multivariate nonlinear non-parametric machine learning for three size fractions, 0.25 μm (panels a–c), 2.5 μm (panels d–f), and 25 μm (g–i).

The left-hand column of plots in Fig. 4 shows the scatter diagrams with the x-axis showing the actual number of particles observed by the in situ optical particle counter and the y-axis showing the number of particles estimated from the NEXRAD data using machine learning. The green circles are the training data, the red pluses are the independent validation dataset, and the blue line shows the ideal response. We can see that for the smaller particles that stay lofted in the air for a long period and do not rapidly sediment, e.g., those with a size of 0.25 μm (Fig. 4a), we have a very good scatter diagram and that the training and independent validation data have almost the same correlation coefficient. The same is true for particles with a diameter of 2.5 μm (Fig. 4d). However, for the larger particles that sediment rapidly, e.g., those with a diameter of 25 μm (Fig. 4g), the independent validation does not do well.

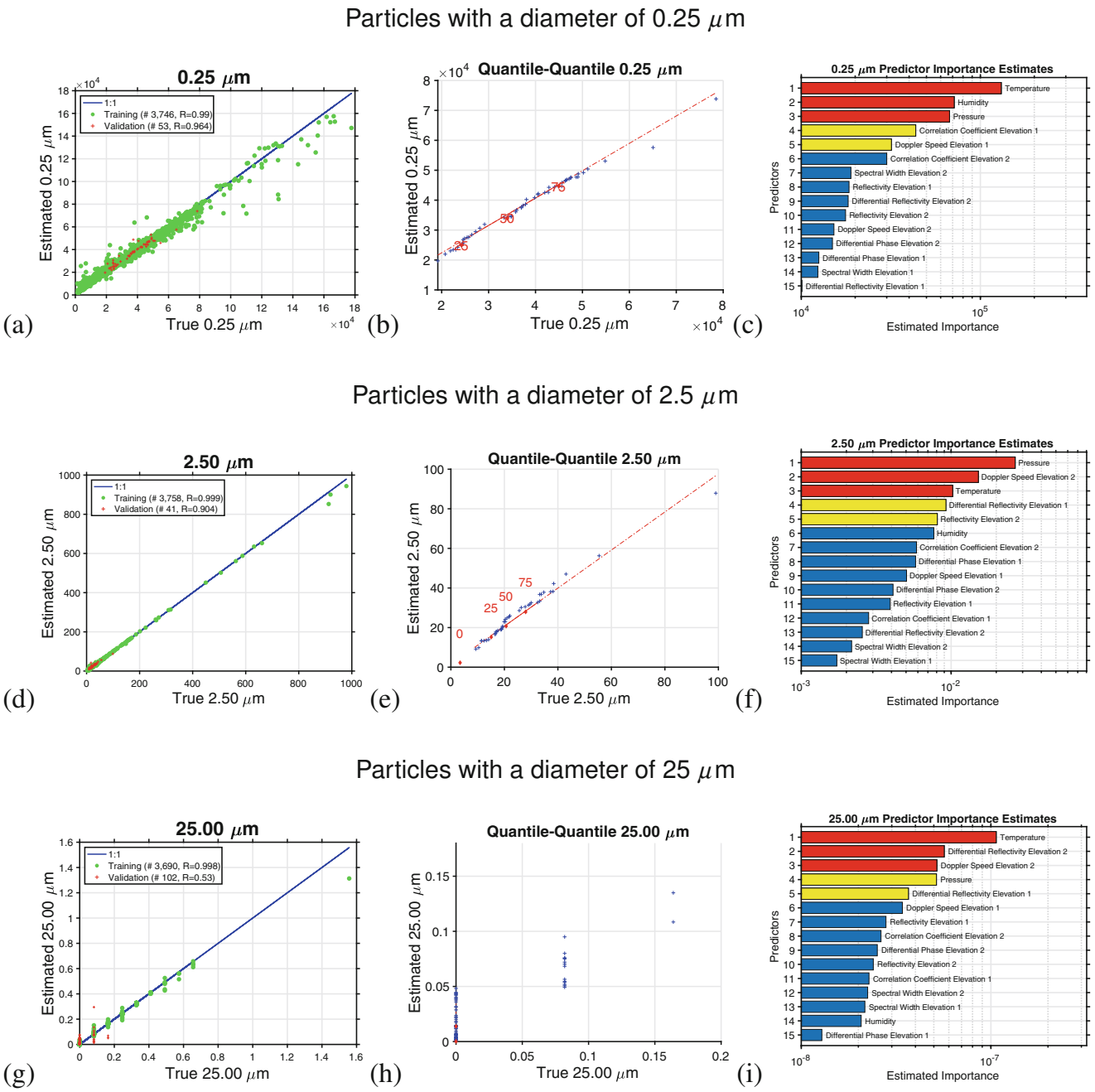


Fig. 4 This figure shows the results of the multivariate nonlinear non-parametric machine learning regression for three size fractions, 0.25 μm (panels **a**–**c**), 2.5 μm (panels **d**–**f**), and 25 μm (**g**–**i**). The left-hand column of plots shows the scatter diagrams with the x-axis showing the actual number of particles observed by the in situ optical particle counter and the y-axis showing the number of particles estimated from the NEXRAD data using machine learning. The green circles are the training data, the red pluses are the independent validation dataset, and the blue line shows the ideal response. The middle column of plots shows the quantile-quantile plots for the machine learning validation data, with the x-axis showing the percentiles from the probability distribution function of the observed number of particles measured by the in situ optical particle counter and the y-axis showing the percentiles from the probability distribution function of the estimated number of particles. The dotted red line shows the ideal response. The right-hand column of plots shows the relative importance of the input variables for estimating the number of particles using machine learning

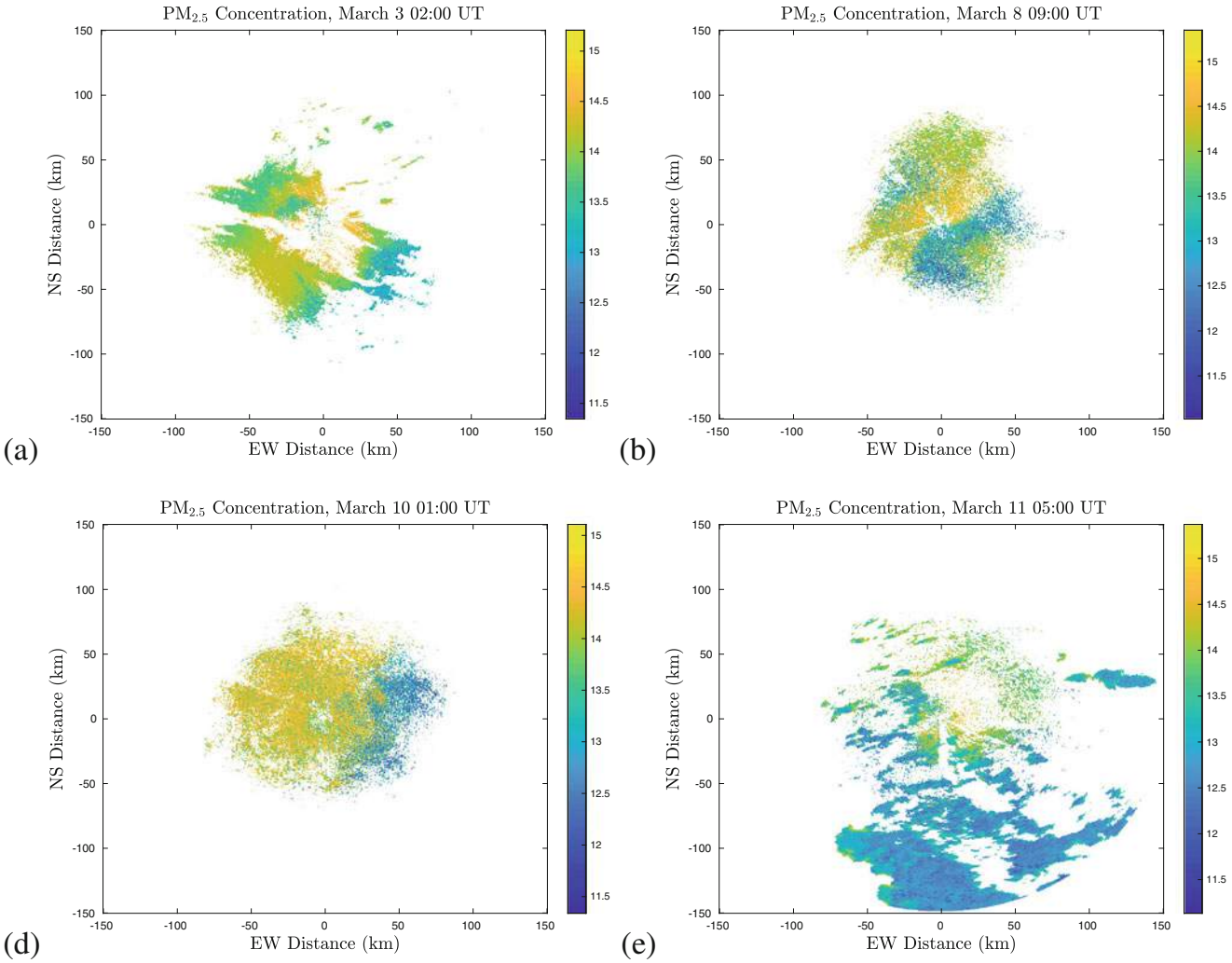


Fig. 5 Showing examples of the distribution of $PM_{2.5}$ over a large spatial area centered at the location of the NEXRAD radar. In this case, the NEXRAD radar measurements over a 0.5 km × 0.5 km are used to estimate the $PM_{2.5}$ concentrations

The middle column of plots in Fig. 4 shows the quantile-quantile plots for the machine learning validation data, with the x-axis showing the percentiles from the probability distribution function of the observed number of particles measured by the in situ optical particle counter and the y-axis showing the percentiles from the probability distribution function (PDF) of the estimated number of particles. The dotted red line shows the ideal response. We can see that for the smaller particles that stay lofted in the air for a long period and do not rapidly sediment, e.g., those with a size of 0.25 μm (Fig. 4b), the shape of the observed and estimated PDFs are almost the same; note that we have a straight line between the 25th and 75th percentiles. The same is true for particles withPlease provide explanation for part labels (a) to (e) in the caption of Fig. 5. a diameter of 2.5 μm (Fig. 4e). However, for the larger particles that sediment rapidly, e.g., those with a diameter of 25 μm (Fig. 4h), the independent validation does not do well.

The right-hand column of plots in Fig. 4 shows the relative importance of the input variables for estimating the number of particles using machine learning. We note that in each case, the temperature and pressure and sometimes the humidity are key factors. For the small particles with a diameter of 0.25 μm, the NEXRAD variables providing the most information are the correlation coefficient and Doppler speed at elevation 1. For the particles with a diameter of 2.5 μm, the NEXRAD variables providing the most information are the Doppler speed at elevation 2 and the differential reflectivity at elevation 1.

Figure 5 shows the spatial distribution of $PM_{2.5}$ particulates estimated over a large spatial area at 0.5 km × 0.5 km resolution. In this case, the machine learning model was developed at 10 km × 10 km pixel, and the model was applied to each pixel using the NEXRAD and atmospheric weather measurements as input.

Summary

Airborne particulates are of particular significance for their human health impacts and their roles in both atmospheric radiative transfer and atmospheric chemistry. Observations of airborne particulates are typically made by environment agencies using rather expensive instruments. Due to the expense of the instruments usually used by environment agencies, the number of sensors that can be deployed is limited. In this study, we have shown two different case studies illustrating the utility of using machine learning for studying airborne particulates.

We have shown that machine learning can be used to effectively calibrate lower-cost optical particle counters. For this calibration, it is critical that measurements of the atmospheric pressure, humidity, and temperature are included. Once the machine learning calibration has been applied to the low-cost sensors, independent validation using scatter diagrams and quantile-quantile plots shows that not only is the calibration effective, but the shape of the resulting probability distribution of observations is very well preserved.

These low-cost sensors are being deployed at scale across the dense urban environment of the Dallas Fort Worth Metroplex for both characterizing the temporal and spatial scales of urban air pollution and providing high schools and high school coaches a tool to assist in making better decisions to reduce adverse health outcomes, e.g., given the levels of pollen/pollution today, should physical education/practice be outside or inside?

In this study, we have also shown that observations made by NEXRAD weather radars can be used with machine learning to effectively estimate the abundance of airborne particulates with a diameter in the size range 0.1–7 μm. For this estimation, it is critical that measurements of the atmospheric pressure, humidity, and temperature are also made. Once machine learning has been applied, scatter diagrams and quantile-quantile plots show that not only is the approach effective, but the shape of the resulting probability distribution of observations is preserved.

References

Ahmad, Z., W. Choi, N. Sharma, J. Zhang, Q. Zhong, D.-Y. Kim, Z. Chen, Y. Zhang, R. Han, D. Shim, et al. 2016. Devices and circuits in CMOS for thz applications. In *Proceedings of the 2016 IEEE International Electron Devices Meeting (IEDM)*, pp. 29–8. New York: IEEE.

Alavi, Amir H., Amir H. Gandomi, and David J. Lary. 2016. *Progress of machine learning in geosciences*.

Albayrak, Arif, J.C. Wei, Maksym Petrenko, D.J. Lary, and G.G. Leptoukh. 2011. Modis aerosol optical depth bias adjustment using machine learning algorithms. In *AGU Fall Meeting Abstracts*.

Alfarra, Mohammedrami. 2004. *Insights into atmospheric organic aerosols using an aerosol mass spectrometer*. PhD thesis, Manchester: University of Manchester.

Allen, Myles R., Vicente R. Barros, John Broome, Wolfgang Cramer, Renate Christ, John A. Church, Leon Clarke, Qin Dahe, Purnamita Dasgupta, Navroz K. Dubash, et al. 2014. *IPCC fifth assessment synthesis report-climate change 2014 synthesis report*.

Alphasense. 2018. *Alphasense user manual opc-n3 optical particle counter*.

Andreae, Meinrat O., and Paul J. Crutzen. 1997. Atmospheric aerosols: Biogeochemical sources and role in atmospheric chemistry. *Science* 276(5315): 1052–1058.

Atkinson, Roger. 2000. Atmospheric chemistry of VOCs and NOx. *Atmospheric Environment* 34(12–14): 2063–2101.

Basu, Mausumi. 2019. The great smog of Delhi. *Lung India: Official Organ of Indian Chest Society* 36(3): 239.

Bickis, Ugis. 1998. Hazard prevention and control in the work environment: airborne dust. *World Health* 13: 16.

Bishop, Christopher M. 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press. 95040465 Christopher M. Bishop. ill.; 24 cm. Includes bibliographical references (p. [457]-475) and index.

Black, J. 2003. Intussusception and the great smog of London, December 1952. *Archives of Disease in Childhood* 88(12): 1040–1042.

Boldo, Elena, Sylvia Medina, Alain Le Tertre, Fintan Hurley, Hans-Guido Mücke, Ferrán Ballester, Inmaculada Aguilera, et al. 2006. Apheis: Health impact assessment of long-term exposure to pm2. 5 in 23 European cities. *European Journal of Epidemiology* 21(6): 449–458.

Bond, Tami C., David G. Streets, Kristen F. Yarber, Sibyl M. Nelson, Jung-Hun Woo, and Zbigniew Klimont. 2004. A technology-based global inventory of black and organic carbon emissions from combustion. *Journal of Geophysical Research: Atmospheres* 109(D14).

Boucher. O. 2015. *Atmospheric Aerosols: Properties and Climate Impacts*. Netherlands: Springer. ISBN 978-9-40-179648-4. https://books.google.co.in/books?id=RXDCoQEACAAJ&redir_esc=y.

Breiman, Leo. 1984. *Classification and Regression Trees*. The Wadsworth Statistics/Probability Series. Belmont: Wadsworth International Group.

Breiman, L. 2001. Random forests. *Machine Learning* 45(1): 5–32.

Broich, Anna V., Lydia E. Gerharz, and Otto Klemm. 2012. Personal monitoring of exposure to particulate matter with a high temporal resolution. *Environmental Science and Pollution Research* 19(7): 2959–2972.

Brown, Molly E., David J. Lary, Anton Vrieling, Demetris Stathakis, and Hamse Mussa. 2008. Neural networks as a tool for constructing continuous NDVI time series from AVHRR and MODIS. *International Journal of Remote Sensing* 29(24): 7141–7158.

Brown, James S., Terry Gordon, Owen Price, and Bahman Asgharian. 2013. Thoracic and respirable particle definitions for human health risk assessment. *Particle and Fibre Toxicology* 10(1): 12.

Byun, Hyang-Min, Tommaso Panni, Valeria Motta, Lifang Hou, Francesco Nordio, Pietro Apostoli, Pier Alberto Bertazzi, and Andrea A Baccarelli. 2013. Effects of airborne pollutants on mitochondrial dna methylation. *Particle and Fibre Toxicology* 10(1): 18.

Carlson, Jen. 2009. Flashback: The city's killer SMOG. https://gothamist.com/news/flashback-the-citys-killer-smog#photo-1.

Carslaw, K.S., R.G. Harrison, and J. Kirkby. 2002. Cosmic rays, clouds, and climate. *Science* 298(5599): 1732–1737.

Chang, Howard H., Anqi Pan, David J. Lary, Lance A. Waller, Lei Zhang, Bruce T. Brackin, Richard W. Finley, and Fazlay S. Faruque. 2019. Time-series analysis of satellite-derived fine particulate matter pollution and asthma morbidity in Jackson, MS. *Environmental Monitoring and Assessment* 191(280).

Charlson, Robert J., S.E. Schwartz, J.M. Hales, Ro D. Cess, Jr J.A. Coakley, J.E. Hansen, and D.J. Hofmann. 1992. Climate forcing by anthropogenic aerosols. *Science* 255(5043): 423–430.

Chen, Renjie, Yi Li, Yanjun Ma, Guowei Pan, Guang Zeng, Xiaohui Xu, Bingheng Chen, and Haidong Kan. 2011. Coarse particles and mortality in three Chinese cities: the china air pollution and health effects study (capes). *Science of the Total Environment* 409(23): 4934–4938.

Cheng, M., and W. Liu. 2009. *Airborne Particulates*. New York: Nova Science Publishers. ISBN 978-1-60-692907-0. https://books.google.co.in/books?id=3H5wPgAACAAJ&redir_esc=y.

Chin, M. 2009. *Atmospheric Aerosol Properties and Climate Impacts*. Collingdale: DIANE Publishing Company. ISBN 978-1-43-791261-6. https://books.google.co.in/books?id=IgJZXXgtHmQC&redir_esc=y.

Choi, Wooyeol, Qian Zhong, Navneet Sharma, Yaming Zhang, Ruonan Han, Z. Ahmad, Dae-Yeon Kim, Sandeep Kshattry, Ivan R. Medvedev, David J. Lary, et al. 2019. Opening terahertz for everyday applications. *IEEE Communications Magazine* 57(8): 70–76.

Chow, Judith C., John G. Watson, et al. 1998. Guideline on speciated particulate monitoring. In *Report prepared for US Environmental Protection Agency, Research Triangle Park, NC*. Reno: Desert Research Institute.

Colbeck, I. 2014. *Aerosol Science: Technology and Applications*. New York: Wiley. ISBN 978-1-11-997792-6. https://books.google.co.in/books?id=eKUTAgAAQBAJ&redir_esc=y.

Colbeck, Ian, and Mihalis Lazaridis. 2010. Aerosols and environmental pollution. *Naturwissenschaften* 97(2): 117–131.

Cortes, C., and V. Vapnik. 1995. Support-vector networks. *Machine Learning* 20(3): 273–297. Times Cited: 3429.

Cromar, Kevin R., Laura A. Gladson, Lars D. Perlmutt, Marya Ghazipura, and Gary W. Ewart. 2016. American thoracic society and marron institute report. Estimated excess morbidity and mortality caused by air pollution above American thoracic society–recommended standards, 2011–2013. *Annals of the American Thoracic Society* 13(8): 1195–1201.

Dadvand, Payam, Jennifer Parker, Michelle L. Bell, Matteo Bonzini, Michael Brauer, Lyndsey A. Darrow, Ulrike Gehring, Svetlana V. Glinianaia, Nelson Gouveia, Eun-hee Ha, et al. 2013. Maternal exposure to particulate air pollution and term birth weight: A multi-country evaluation of effect and heterogeneity. *Environmental Health Perspectives* 121(3): 267.

Demuth, Howard B., Mark H. Beale, Orlando De Jess, and Martin T. Hagan. 2014. *Neural Network Design*. Martin Hagan, USA, 2nd edn. ISBN 0-9717-3211-6, 978-0-97-173211-7.

Dockery, D.W., C.A. Pope, X.P. Xu, J.D. Spengler, J.H. Ware, M.E. Fay, et al. 1993a. An association between air-pollution and mortality in 6 United-States cities. *New England Journal of Medicine* 329(24): 1753–1759. *Find this article online*.

Dockery, Douglas W., C. Arden Pope, Xiping Xu, John D. Spengler, James H. Ware, Martha E. Fay, Benjamin G. Ferris Jr, and Frank E. Speizer. 1993b. An association between air pollution and mortality in six US cities. *New England Journal of Medicine* 329(24): 1753–1759.

Domingos, Pedro. 2015. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. London: Basic Books.

Dong, W., G. Guan, Y. Chen, K. Guo, and Y. Gao. 2015. Mosaic: Towards city scale sensing with mobile sensor networks. In *Proceedings of the 2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS)*, pp. 29–36. https://doi.org/10.1109/ICPADS.2015.12.

Dubovik, Oleg, Brent Holben, Thomas F. Eck, Alexander Smirnov, Yoram J. Kaufman, Michael D. King, Didier Tanré, and Ilya Slutsker. 2002. Variability of absorption and optical properties of key aerosol types observed in worldwide locations. *Journal of the Atmospheric Sciences* 59(3): 590–608.

Gali, Rohith Kumar. 2010. *Assessment of NEXRAD P3 data on streamflow simulation using SWAT for North Fork Ninnescah Watershed, Kansas*. PhD thesis, Manhattan: Kansas State University.

Glasser, Marvin, Leonard Greenburg, and Franklyn Field. 1967. Mortality and morbidity during a period of high levels of air pollution: New York, Nov 23–25, 1966. *Archives of Environmental Health: An International Journal* 15(6): 684–694.

Guenther, A., T. Karl, Pedro Harley, C. Wiedinmyer, P.I. Palmer, and C. Geron. 2006. Estimates of global terrestrial isoprene emissions using MEGAN (model of emissions of gases and aerosols from nature). *Atmospheric Chemistry and Physics* 6(11): 3181–3210.

Guo, Liqiong, Hyang-Min Byun, Jia Zhong, Valeria Motta, Jitendra Barupal, Yinan Zheng, Chang Dou, Feiruo Zhang, John P McCracken, Anaité Diaz, et al. 2014. Effects of short-term exposure to inhalable particulate matter on DNA methylation of tandem repeats. *Environmental and Molecular Mutagenesis* 55(4): 322–335.

Haberzettl, Petra, Timothy E. O'Toole, Aruni Bhatnagar, and Daniel J. Conklin. 2016. Exposure to fine particulate air pollution causes vascular insulin resistance by inducing pulmonary oxidative stress. *Environmental Health Perspectives* 124(12): 1830.

Hallquist, Mattias, John C. Wenger, Urs Baltensperger, Yinon Rudich, David Simpson, M. Claeys, J. Dommen, N.M. Donahue, C. George, A.H. Goldstein, et al. 2009. The formation, properties and impact of secondary organic aerosol: Current and emerging issues. *Atmospheric Chemistry and Physics* 9(14): 5155–5236.

Harrison, William Alan. 2015. *In-situ observation of atmospheric particulates*. Dallas: The University of Texas.

Harrison, William A., David Lary, Brian Nathan, and Alec G Moore. 2015. The neighborhood scale variability of airborne particulates. *Journal of Environmental Protection* 6(05): 464.

Haykin, Simon S. 1994. *Neural Networks: A Comprehensive Foundation*. New York: Macmillan. 93028092 Simon Haykin. ill.; 26 cm. Includes bibliographical references (p. 635–690) and index.

Haykin, Simon S. 1999. *Neural Networks: A Comprehensive Foundation*. Upper Saddle River: Prentice Hall, 2nd edn., 98007011 Simon Haykin. ill.; 25 cm. Includes bibliographical references (p. 796–836) and index.

Haykin, Simon S. 2001. Kalman Filtering and Neural Networks. In *Adaptive and Learning Systems for Signal Processing, Communications, and Control*. New York: Wiley. 2001049240 edited by Simon Haykin. ill.; 24 cm. A Wiley Interscience publication. Includes bibliographical references and index.

Haykin, Simon S. 2007. *New Directions in Statistical Signal Processing: From Systems to Brain*. Neural Information Processing Series. Cambridge: MIT Press. 2005056210 GBA671791 013536699 (OCoLC)ocm62302576 (OCoLC)62302576 edited by Simon Haykin ... [et al.]. ill.; 26 cm. Includes bibliographical references (p. [465]-508) and index. Modeling the mind: from circuits to systems/Suzanna Becker—Empirical statistics and stochastic models for visual signals/David Mumford—The machine cocktail party problem/Simon Haykin, Zhe Chen—Sensor adaptive signal processing of biological nanotubes (ion channels) at macroscopic and nano scales/Vikram Krishnamurthy—Spin diffusion: a new perspective in magnetic resonance imaging/Timothy R. Field—What makes a dynamical system computationally powerful?/Robert Legenstein, Wolfgang Maass—A variational principle for graphical models/Martin J. Wainwright, Michael I. Jordan—Modeling large dynamical systems with dynamical consistent neural networks/Hans-Georg Zimmermann ... [et al.]—Diversity in communication: from source coding to wireless networks/Suhas N. Diggavi—Designing patterns for easy recognition: information transmission with low-density parity-check codes/Frank R. Kschischang, Masoud Ardakani—Turbo processing/Claude Berrou, Charlotte Langlais, Fabrice Seguin—Blind

signal processing based on data geometric properties/Konstantinos Diamantaras—Game-theoretic learning / Geoffrey J. Gordon—Learning observable operator models via the efficient sharpening algorithm/Herbert Jaeger . . . [et al.].

Health Effects Institute HEI. 2017. *State of global air 2017*. Special.

Hinds, William C. 2012. *Aerosol technology: properties, behavior, and measurement of airborne particles*. New York: Wiley.

Ho, T.K. 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(8): 832–844.

Holben, B.N., D. Tanre, A. Smirnov, T.F. Eck, I. Slutsker, N. Abuhassan, W.W. Newcomb, J.S. Schafer, B. Chatenet, F. Lavenu, et al. 2001. An emerging ground-based aerosol climatology: Aerosol optical depth from AERONET. *Journal of Geophysical Research: Atmospheres* 106(D11): 12067–12097.

Huang, Fang, Renjie Chen, Yuetian Shen, Haidong Kan, and Xingya Kuang. 2016. The impact of the 2013 eastern china smog on outpatient visits for coronary heart disease in shanghai, china. *International Journal of Environmental Research and Public Health* 13(7): 627.

Jimenez, Jose L. M.R. Canagaratna, N.M. Donahue, A.S.H. Prevot, Qi Zhang, Jesse H. Kroll, Peter F. DeCarlo, James D. Allan, H. Coe, N.L. Ng, et al. 2009. Evolution of organic aerosols in the atmosphere. *Science* 326(5959): 1525–1529.

Kanakidou, M. J.H. Seinfeld, S.N. Pandis, I. Barnes, F.J. Dentener, M.C. Facchini, R. Van Dingenen, B. Ervens, ANCJSE Nenes, C.J. Nielsen, et al. 2005. Organic aerosol and global climate modelling: a review. *Atmospheric Chemistry and Physics* 5(4): 1053–1123.

Kelly, Frank J., and Julia C Fussell. 2016. Health effects of airborne particles in relation to composition, size and source. *Airborne Particulate Matter*, 344–382.

Kirkby, Jasper, Joachim Curtius, João Almeida, Eimear Dunne, Jonathan Duplissy, Sebastian Ehrhart, Alessandro Franchin, Stéphanie Gagné, Luisa Ickes, Andreas Kürten, et al. 2011. Role of sulphuric acid, ammonia and galactic cosmic rays in atmospheric aerosol nucleation. *Nature* 476(7361): 429.

Kneen, Melanie A., David J. Lary, William A. Harrison, Harold J. Annegarn, and Tom H. Brikowski. 2016. Interpretation of satellite retrievals of pm2.5 over the Southern African interior. *Atmospheric Environment* 128: 53–64.

Kokhanovsky, Alexander A. 2008. *Aerosol optics: light absorption and scattering by particles in the atmosphere*. Berlin: Springer.

Kondratyev, Kirill Ya, Lev S. Ivlev, Vladimir F. Krapivin, and Costas A. Varostos. 2006. *Atmospheric aerosol properties: Formation, processes and impacts*. Berlin: Springer.

Lary, D. 2007. Using neural networks for instrument cross-calibration. In *AGU Fall Meeting Abstracts*.

Lary, David John. 2010. *Artificial intelligence in geoscience and remote sensing*. London: INTECH Open Access Publisher.

Lary, David J. 2013. Using multiple big datasets and machine learning to produce a new global particulate dataset: A technology challenge case study. In *AGU Fall Meeting Abstracts*.

Lary, David John. 2014. Bigdata and machine learning for public health. In *142nd APHA Annual Meeting and Exposition 2014*. Washington: APHA.

Lary, D.J. and O. Aulov. 2008. Space-based measurements of hcl: Intercomparison and historical context. *Journal of Geophysical Research: Atmospheres* 113(D15).

Lary, D.J., M.D. Müller, and H.Y. Mussa. 2003. Using neural networks to describe tracer correlations. *Atmospheric Chemistry and Physics Discussions* 3(6): 5711–5724.

Lary, D.J., M.D. Müller, and H.Y. Mussa. 2004. Using neural networks to describe tracer correlations. *Atmospheric Chemistry and Physics* 4(1): 143–146.

Lary, David J., L.A. Remer, Devon MacNeill, Bryan Roscoe, and Susan Paradise. 2009a. Machine learning and bias correction of MODIS aerosol optical depth. *IEEE Geoscience and Remote Sensing Letters* 6(4): 694–698.

Lary, D.J., L.A. Remer, D. MacNeill, B. Roscoe, and S. Paradise. 2009b. Machine learning and bias correction of MODIS aerosol optical depth. *IEEE Geoscience and Remote Sensing Letters* 6(4): 694–698.

Lary, D.J., A. Nikitkov, D. Stone, and Alexey Nikitkov. 2010. Which machine-learning models best predict online auction seller deception risk. *American Accounting Association AAA Strategic and Emerging Technologies*.

Lary, David J., Fazlay S. Faruque, Nabin Malakar, Alex Moore, Bryan Roscoe, Zachary L. Adams, and York Eggelston. 2014. Estimating the global abundance of ground level presence of particulate matter (pm2. 5). *Geospatial Health* 8(3): 611–630.

Lary, D.J., T. Lary, and B. Sattler. 2015a. Using machine learning to estimate global pm2.5 for environmental health studies. *Environmental Health Insights* 9: EHI–S15664.

Lary, D.J., T. Lary, and B. Sattler. 2015b. Using machine learning to estimate global pm2. 5 for environmental health studies. *Environmental Health Insights* 9(Suppl 1): 41.

Lary, David J., Amir H. Alavi, Amir H. Gandomi, and Annette L. Walker. 2016. Machine learning in geosciences and remote sensing. *Geoscience Frontiers* 7(1): 3–10.

Lary, David J., Gebreab K. Zewdie, Xun Liu, Daji Wu, Estelle Levetin, Rebecca J. Allee, Nabin Malakar, Annette Walker, Hamse Mussa, Antonio Mannino, et al. 2018. Machine learning applications for earth observation. In *Earth Observation Open Science and Innovation. ISSI Scientific Report Series* vol. 15, pp. 165–218. Berlin: Springer.

Lary, Maria-Anna, Leslie Allsop, and David John Lary. 2019. Using machine learning to examine the relationship between asthma and absenteeism. *Environmental Modeling and Assessment* 191(332): 1–9.

Lelieveld, Jos, John S. Evans, M. Fnais, Despina Giannadaki, and Andrea Pozzer. 2015. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature* 525(7569): 367.

Lesins, Glen, Petr Chylek, and Ulrike Lohmann. 2002. A study of internal and external mixing scenarios and its effect on aerosol optical properties and direct radiative forcing. *Journal of Geophysical Research: Atmospheres* 107(D10): AAC–5.

Levy, Robert. 2014. Smog shrouds Eastern China. https://earthobservatory.nasa.gov/images/82535/smog-shrouds-eastern-china.

Li, Linglong, Yixin Zheng, and Lin Zhang. 2014. Demonstration abstract: Pimi air box: a cost-effective sensor for participatory indoor quality monitoring. In *Proceedings of the 13th International Symposium on Information Processing in Sensor Networks*, pp. 327–328. New York: IEEE Press.

Lim, Stephen S., Theo Vos, Abraham D. Flaxman, Goodarz Danaei, Kenji Shibuya, Heather Adair-Rohani, Mohammad A. AlMazroa, Markus Amann, H. Ross Anderson, Kathryn G. Andrews, et al. 2012. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: A systematic analysis for the global burden of disease study 2010. *The Lancet* 380(9859): 2224–2260.

Madonna, F., A. Amodeo, G. D'Amico, L. Mona, and G. Pappalardo. 2010. Observation of non-spherical ultragiant aerosol using a microwave radar. *Geophysical Research Letters* 37(21).

Maji, Kamal Jyoti, Anil Kumar Dikshit, and Ashok Deshpande. 2017. Disability-adjusted life years and economic cost assessment of the health effects related to pm2. 5 and pm10 pollution in Mumbai and Delhi, in India from 1991 to 2015. *Environmental Science and Pollution Research* 24(5): 4709–4730.

Malakar, Nabin K., David J. Lary, A. Moore, D. Gencaga, Bryan Roscoe, Arif Albayrak, and Jennifer Wei. 2012a. Estimation and bias correc-

tion of aerosol abundance using data-driven machine learning and remote sensing. In *Proceedings of the2012 Conference on Intelligent Data Understanding*, pp. 24–30. New York: IEEE.

Malakar, N.K., D.J. Lary, R. Allee, R. Gould, and D. Ko. 2012b. Towards automated ecosystem-based management: A case study of northern Gulf of Mexico water. In *AGU Fall Meeting Abstracts*.

Malakar, N.K., D.J. Lary, D. Gencaga, A. Albayrak, and J. Wei. 2013. Towards identification of relevant variables in the observed aerosol optical depth bias between MODIS and Aeronet observations. In *AIP Conference Proceedings*, vol. 1553, pp. 69–76. College Park: AIP.

Malakar, Nabin K., D.J. Lary, and B. Gross. 2018. Case studies of applying machine learning to physical observation. In *AGU Fall Meeting Abstracts*.

Mannucci, Pier Mannuccio. 2017. *Air pollution levels and cardiovascular health: Low is not enough*.

McCulloch, W.S., and W. Pitts. 1943. *Bulletin of Mathematical Biophysics* 5: 115. https://doi.org/10.1007/BF02478259.

Medvedev, Ivan R., Robert Schueler, Jessica Thomas, O. Kenneth, Hyun-Joo Nam, Navneet Sharma, Qian Zhong, David J. Lary, and Philip Raskin. 2016. Analysis of exhaled human breath via terahertz molecular spectroscopy. In *Proceedings of the 2016 41st International Conference on Infrared, Millimeter, and Terahertz waves (IRMMW-THz)*, pp. 1–2. New York: IEEE.

Nada Osseiran, Lindmeier, Christian. 2018. 9 out of 10 people worldwide breathe polluted air, but more countries are taking action. https://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action.

Nathan, Brian J., and David J. Lary. 2019. Combining domain filling with a self-organizing map to analyze multi-species hydrocarbon signatures on a regional scale. *Environmental Modeling and Assessment* 191(337).

O, K.K., Q. Zhong, N. Sharma, W. Choi, R. Schueler, I.R. Medvedev, H.-J. Nam, P. Raskin, F.C. De Lucia, J.P. McMillan, et al. 2017. Demonstration of breath analyses using CMOS integrated circuits for rotational spectroscopy. In *International Workshop on Nanodevice Technologies, Hiroshima, Japan*.

Oberdörster, Günter, Eva Oberdörster, and Jan Oberdörster. 2005. Nanotoxicology: An emerging discipline evolving from studies of ultrafine particles. *Environmental Health Perspectives* 113(7): 823–839.

Onishi, Kazunari, Tsuyoshi Thomas Sekiyama, Masanori Nojima, Yasunori Kurosaki, Yusuke Fujitani, Shinji Otani, Takashi Maki, Masato Shinoda, Youichi Kurozawa, and Zentaro Yamagata. 2018. Prediction of health effects of cross-border atmospheric pollutants using an aerosol forecast model. *Environment International* 117: 48–56.

Pascal, Mathilde, Magali Corso, Olivier Chanel, Christophe Declercq, Chiara Badaloni, Giulia Cesaroni, Susann Henschel, Kadri Meister, Daniela Haluza, Piedad Martin-Olmedo, et al. 2013. Assessing the public health impacts of urban air pollution in 25 european cities: results of the Aphekom project. *Science of the Total Environment* 449: 390–400.

Polivka, Barbara J. The great London smog of 1952. *AJN The American Journal of Nursing* 118(4): 57–61 (2018).

Pope, C. Arden, Richard T. Burnett, George D. Thurston, Michael J. Thun, Eugenia E. Calle, Daniel Krewski, and John J. Godleski. 2004. Cardiovascular mortality and long-term exposure to particulate air pollution. *Circulation* 109(1): 71–77.

Pope, C., Arden Dockery, and Douglas W. 2006. Health effects of fine particulate air pollution: Lines that connect. *Journal of the Air and Waste Management Association* 56(6): 709–742.

Pope C. ArdenIII, Richard T. Burnett, Michael J. Thun, Eugenia E. Calle, Daniel Krewski, Kazuhiko Ito, and George D. Thurston. 2002. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Jama* 287(9): 1132–1141.

Pöschl. Ulrich. 2005. Atmospheric aerosols: composition, transformation, climate and health effects. *Angewandte Chemie International Edition* 44(46): 7520–7540.

Pun, Vivian C., Justin Manjourides, and Helen Suh. 2017. Association of ambient air pollution with depressive and anxiety symptoms in older adults: results from the NSHAP study. *Environmental Health Perspectives* 125(3): 342.

Ramanathan, V.C.P.J., P.J. Crutzen, J.T. Kiehl, and Dm Rosenfeld. 2001. Aerosols, climate, and the hydrological cycle. *Science* 294(5549): 2119–2124.

Ruckerl, R., A. Ibald-Mulli, W. Koenig, A. Schneider, G. Woelke, J. Cyrys, and A. Peters. 2006. Air pollution and markers of inflammation and coagulation in patients with coronary heart disease. *American Journal of Respiratory and Critical Care Medicine* 173(4): 432–441.

Safavian, S.R., and D. Landgrebe. 1991. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics* 21(3): 660–674. ISSN 0018–9472. https://doi.org/10.1109/21.97458.

Santibañez, Daniela A., Sergio Ibarra, Patricia Matus, Rodrigo Seguel, et al. 2013. A five-year study of particulate matter (pm2. 5) and cerebrovascular diseases. *Environmental Pollution* 181: 1–6.

Saravanan, J., M. Jayadurgalakshmi, and R. Karthickraja. 2017. China's Nanjing vs India's Delhi–a perspective for vertical forest. *International Journal of Civil Engineering and Technology* 8: 12.

Schauer, James J., Wolfgang F. Rogge, Lynn M. Hildemann, Monica A. Mazurek, Glen R. Cass, and Bernd R.T. Simoneit. 1996. Source apportionment of airborne particulate matter using organic compounds as tracers. *Atmospheric Environment* 30(22): 3837–3855.

Seinfeld, J.H. 1986. *Atmospheric chemistry and physics of air pollution*. A Wiley-Interscience publication. New York: Wiley. ISBN 978-0-47-182857-0. https://books.google.co.in/books?id=NAhSAAAAMAAJ&redir_esc=y.

Shy, Carl M., Victor Hasselblad, Robert M. Burt, Cornelius J. Nelson, and Arlan A. Cohen. 1973. Air pollution effects on ventilatory function of us schoolchildren: Results of studies Cincinnati, Chattanooga, and New York. *Archives of Environmental Health: An International Journal* 27(3): 124–128.

Solomon, Feliz. 2016. China's SMOG is as deadly as smoking, new research claims. https://time.com/4617295/china-smog-smoking-environment-air-pollution/.

Spira-Cohen, Ariel, Lung Chi Chen, Michaela Kendall, Ramona Lall, and George D. Thurston. 2011. Personal exposures to traffic-related air pollution and acute respiratory health among Bronx schoolchildren with asthma. *Environmental Health Perspectives* 119(4): 559.

Stier, P., J. Feichter, S. Kinne, S. Kloster, E. Vignati, J. Wilson, L. Ganzeveld, I. Tegen, Martin Werner, Y. Balkanski, et al. 2005. The aerosol-climate model echam5-ham. *Atmospheric Chemistry and Physics* 5(4): 1125–1156.

Stocker, Thomas. 2014. *Climate change 2013: The physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Streets, D. Ga, T.C. Bond, G.R. Carmichael, S.D. Fernandes, Q. Fu, D. He, Z. Klimont, S.M. Nelson, N.Y. Tsai, M. Qm Wang, et al. 2003. An inventory of gaseous and primary aerosol emissions in asia in the year 2000. *Journal of Geophysical Research: Atmospheres* 108(D21).

Sỳkorová, Barbora, Marek Kucbel, and Konstantin Raclavskỳ. 2016. Composition of airborne particulate matter in the industrial area versus mountain area. *Perspectives in Science* 7: 369–372.

Terry, James P., Gensuo Jia, Robert Boldi, and Sarah Khan. 2018. The Delhi 'gas chamber': smog, air pollution and the health emergency of november 2017. *Weather* 73(11): 348–352.

Thurston, George D., Jiyoung Ahn, Kevin R Cromar, Yongzhao Shao, Harmony R. Reynolds, Michael Jerrett, Chris C. Lim, Ryan Shanley, Yikyung Park, and Richard B. Hayes. 2016. Ambient particulate matter air pollution exposure and mortality in the nih-aarp diet and health cohort. *Environmental Health Perspectives* 124(4): 484.

US EPA. 2004. *Air quality criteria for particulate matter*, vol. 2. US Environmental Protection Agency, Office of Research and Development, National Center for Environmental Assessment.

Vapnik, Vladimir Naumovich. 1982. *Estimation of Dependences Based on Empirical Data*. In *Springer Series in Statistics*. New York: Springer.

Vapnik, Vladimir Naumovich. 1995. *The Nature of Statistical Learning Theory*. New York: Springer.

Vapnik, Vladimir Naumovich. 2000. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science, 2nd edn. Springer, New York.

Vapnik, Vladimir Naumovich. 2006. *Estimation of Dependences Based on Empirical Data; Empirical Inference Science: Afterword of 2006*. In *Information Science and Statistics*, 2nd edn. New York: Springer.

Wachs, Anthony. 2009. A dem-dlm/fd method for direct numerical simulation of particulate flows: Sedimentation of polygonal isometric particles in a Newtonian fluid with collisions. *Computers and Fluids* 38(8): 1608–1628.

Wilkins, E.T. 1954. Air pollution and the London fog of December 1952. *Journal of the Royal Sanitary Institute* 74(1): 1–21.

Wu, Daji, Gebreab K. Zewdie, Xun Liu, Melanie Anne Kneen, and David John Lary. 2017. Insights into the morphology of the East Asia pm2. 5 annual cycle provided by machine learning. *Environmental Health Insights* 11: 1178630217699611.

Wu, Daji, David J. Lary, Gebreab K. Zewdie, and Xun Liu. 2019. Using machine learning to understand the temporal morphology of the pm2.5 annual cycle in East Asia. *Environmental Monitoring and Assessment* 191(272): 1–14.

Yunker, Mark B., Robie W. Macdonald, Roxanne Vingarzan, Reginald H. Mitchell, Darcy Goyette, and Stephanie Sylvestre. 2002. PAHs in the Fraser river basin: a critical appraisal of PAH ratios as indicators of PAH source and composition. *Organic Geochemistry* 33(4): 489–515.

Zewdie, Gebreab, and David J. Lary. 2018. Applying machine learning to estimate allergic pollen using environmental, land surface and NEXRAD radar parameters. In *AGU Fall Meeting Abstracts*.

Zewdie, Gebreab K., David J. Lary, Estelle Levetin, and Gemechu F. Garuma. 2019a. Applying deep neural networks and ensemble machine learning methods to forecast airborne ambrosia pollen. *International Journal of Environmental Research and Public Health* 16(11): 1992.

Zewdie, Gebreab K., David J. Lary, Xun Liu, Daji Wu, and Estelle Levetin. 2019b. Estimating the daily pollen concentration in the atmosphere using machine learning and NEXRAD weather radar data. *Environmental Monitoring and Assessment* 191(7): 418.

 doi:10.20944/preprints202203.0200.v1

Article

Unsupervised Blink Detection Using Eye Aspect Ratio Values

Bharana Fernando ‡, Arjun Sridhar ‡, Shawhin Talebi ‡, John Waczak, and David J. Lary ‡

Hanson Center for Space Sciences, University of Texas at Dallas, Richardson TX 75080, USA
* Correspondence: aaf170130@utdallas.edu
‡ These authors contributed equally to this work.

Abstract: The eyes serve as a window into underlying physical and cognitive processes. Although factors such as pupil size have been studied extensively, a less explored yet potentially informative aspect is blinking. Given its novelty, blink detection techniques are far less available compared to eye-tracking and pupil size estimation tools. In this work, we present a new unsupervised machine learning blink detection strategy using existing eye-tracking technology. The method is compared to two existing techniques. All three algorithms make use of eye aspect ratio values for blink detection. Accurate and rapid blink detection complements existing eye-tracking research and may provide a new informative index of physical and mental status.

Keywords: Machine Learning; Eye Tracking; Blink Detection

1. Introduction

It has been said that the eyes are a “window to the soul". Despite its colloquial nature, such a statement has an anatomical basis. The eyes play a key role in the central nervous system. For example, pupils dilate in response to elevated levels of autonomic arousal and mental effort, gaze converges and diverges based on attention [1–4]. Although these phenomena have been studied extensively, there is another aspect of the eyes that perhaps deserves more attention: blinking.

Human adults typically blink 15-20 times a minute. This action is physiologically necessary to keep the eye lubricated, but that is only required about two to four times a minute. On average, the duration of a blink can range from 100-400 ms, but it also depends on individual characteristics, fatigue level, and the time of day [5].

Effective classification of blinks has a wide variety of applications. For instance, it has been used to assess an individual’s mental status. This includes the evaluation of fatigue, concentration levels, attention span, and cognitive load [6,7]. Another application of blink detection is to facilitate the removal of blinking artifacts from Electroencephalography (EEG) signals [8].

Although blinking can be controlled directly, it is often involuntary. Thus, autonomically regulated blink rates may be indicative of different cognitive states. When a person relaxes or concentrates on a visual object, their blinking rate decreases, whereas negative emotions and conversations with other people cause it to increase [5].

There are three observed types of blinks: spontaneous, reflex, and voluntary. The first two types of blinks are autonomic responses. Spontaneous blinks occur without external stimuli, while reflexive blinks depend on bright lights, loud noises, etc. Furthermore, the closing phase of the spontaneous blink is longer than the reflex blink, while the opposite is true for the opening phase, indicating that the eye seeks to autonomously protect itself at a moment’s notice [9]. It has been found that the spontaneous eye blink rate correlates with dopaminergic activity and striatal dopamine receptor availability [10,11]. Many studies have shown that dopamine plays a role in attention [12], learning [13], goal-directed behavior [14], and time perception [15]. A quantification of dopamine can provide insight into these critical cognitive functions. Considering that spontaneous blinks are correlated with dopaminergic activity, they may serve as a convenient proxy.

In this paper, we compare two existing blink classifiers to a novel blink detection technique that employs an unsupervised machine learning method. We find that our

Citation: Fernando, B.; Sridhar, A.; Talebi, S. Title. *Preprints* **2021**, *1*, 0. https://doi.org/

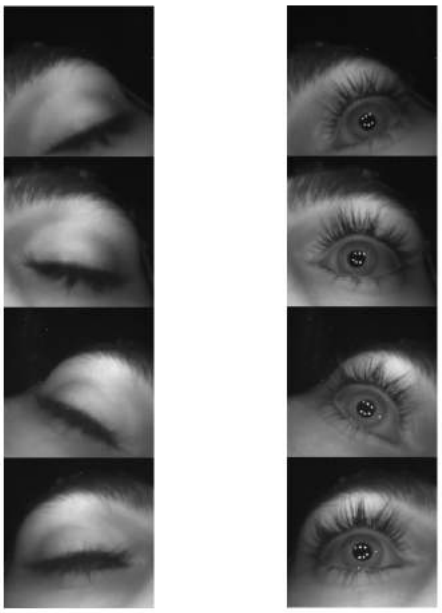
Received:
Accepted:
Published:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2022 doi:10.20944/preprints202203.0200.v1

unsupervised method outperforms an existing supervised machine learning method, thus eliminating the need the need for a labeled training dataset.

2. Materials and Methods

Eye data were collected using the Tobii Pro Glasses 2 eye tracking system. In addition to gaze and pupil tracking, the glasses record infrared images of the left and right eyes from two angles, at a rate of 50 Hz. These images are stacked into a single frame and saved as a video. We will refer to this video as an eye-stream video. Example frames from an eye-stream video can be seen in Figure 1.



1.png

Figure 1. Example video frames from eye-stream video showing eyes-closed and eyes-open frames.

Although the Tobii Pro Glasses 2 collect over 30 biometric variables at 100Hz, it lacks a built-in blink detection feature. This is a major setback for a wide range of biometrics research toward attention, cognitive load, and dopamine, in which blinking may play a central role. Given the information rich eye-stream video, however, we demonstrate that reliable blink detection can be achieved.

2.1. Eye Aspect Ratio (EAR)

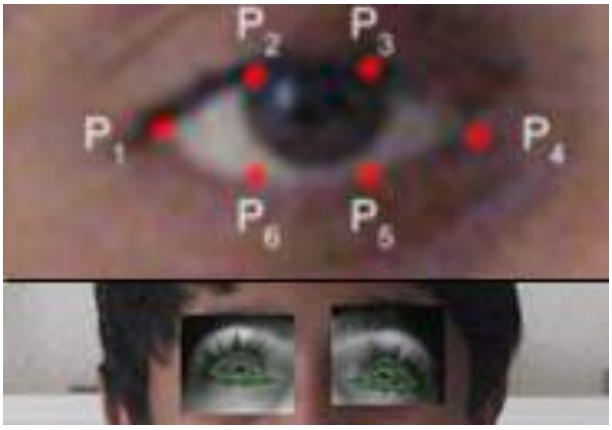
The eye aspect ratio (EAR) is the key concept on which all presented algorithms are built. An EAR value is a numerical representation of how open or closed the eye is [16]. The defining equation is a ratio between the height and width of the eye as calculated by:

$$EAR = \frac{||p2 - p6|| + ||p3 - p5||}{||p1 - p4||}, \tag{1}$$

Where, $p1$ through $p6$ are points contouring the eye, as seen on Figure 2. As the eye closes, the numerator (or height) will approach zero. Thus, a low EAR value may represent a blink occurring. The six points in the equation are facial landmarks automatically detected using the method from [17] as shown in the top panel of Figure 2.

Using the eye landmarks and equation 1, we can calculate the EAR value for each eye, and subsequently an average EAR value is considered, as eye blinking is synchronous [16]. However, the facial landmark detection method described above requires the entire face to be captured, but the Tobii Pro Glasses 2 only captures eye images as shown in Figure 1. To overcome this, the eyes were cropped from the eye-stream video and then superimposed onto a face image. Right panel of Figure 2 shows the result. Using the structure in Figure 2

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2022 doi:10.20944/preprints202203.0200.v1



combined.png

Figure 2. The top panel shows labeled eye landmarks that equation refeq1 utilizes. The bottom panel shows a cropped image of the eyes from the Tobii Pro glasses superimposed on a face image so that the eye landmarks can be detected.

and the EAR value equation, the algorithms discussed in the following sections attempt to detect blinks.

2.2. Baseline Method

The baseline method is adapted from [16], using the EAR values obtained from the previous section. This algorithm detects blinks by altering two parameters: a threshold EAR value and a number of consecutive frames below said threshold. If the average EAR value falls below the threshold for some number of consecutive frames, we classify this as a blink [16].

When the eyes are open, the average EAR value will remain relatively constant, but as they begin to close there is a sharp drop. Figure 3 shows the plot of the EAR values for each frame from the eye-stream video with a threshold of 0.2.

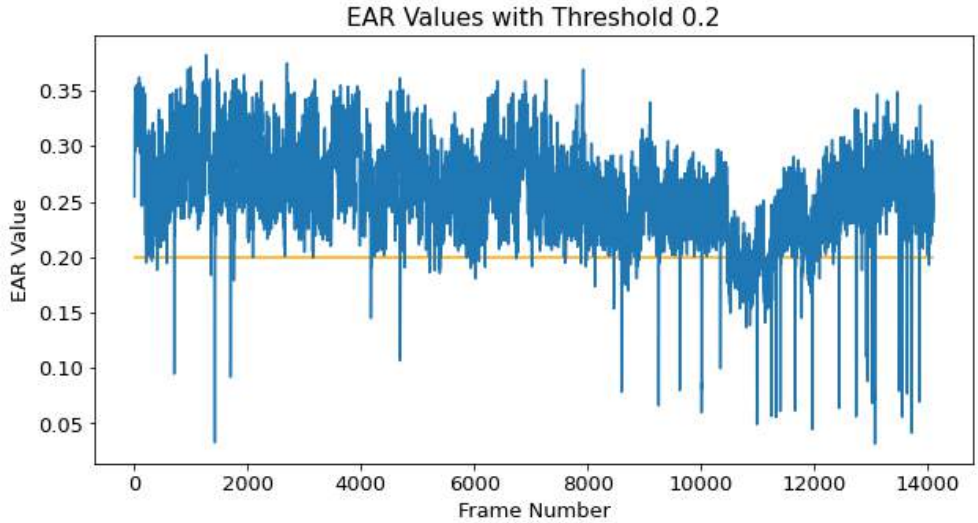


Figure 3. The plot of EAR Values for each frame using the structure in Figure 2 with a threshold of 0.2. A blink is detected if the EAR value is below 0.2 for 3 consecutive frames.

From Figure 3, an EAR value of 0.2 appears to be a reasonable threshold, and 3 consecutive frames were chosen to detect a blink. Although thresholding is a natural attempt to classify blinks, it is not robust to noise. As can be seen from Figure 3, EAR values are very noisy. Additionally, superimposing the eyes from the eye-stream video further amplifies noise when it comes to detecting the landmarks.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2022 doi:10.20944/preprints202203.0200.v1

This leads to several false positives in blink detection. EAR values may fall below the threshold for a consecutive number of frames even when the person is not blinking. Conversely, lowering the threshold produces true negatives, or missing out on blinks. Therefore, this method requires sophistication in terms of choosing a robust threshold and consecutive number of frames. The remaining two methods attempt to overcome this issue by using machine learning.

2.3. Supervised Learning Method (Support Vector Machines)

In this section, we present method offered in [16] which uses a supervised machine learning method, the support vector machine (SVM). It is a way of improving blink classification as it helps combat against the noise that is generated from the baseline method as described in the previous section.

The idea behind SVM is as follows: given a set of data points that are in different classes and represented in an n-dimensional space, the model will find the best line that separates the points into their respective classes. Thus, SVM is a good choice for dealing with noisy and high dimensional data, as it is not affected by local minima (e.g. if a person is squinting, the baseline method may categorize this as a blink) [16]. For more details on SVM we refer the reader to [18].

To implement this method, it is proposed that the SVM model use a 13 dimension feature vector, where each vector consists of the average EAR value for the current frame, the previous 6 frames, and the next 6 frames (as shown in Figure 4) [16].

5.png 

Figure 4. The 13 dimension average EAR value window shown for frame N.

This 13 frame feature window allows for a larger temporal period, where noise can be accounted for, which results in significantly improved blink classification.

For our model, we used a dataset consisting of 9,042 static eye-stream images. The average EAR value was computed for each image, and then labeled as open or closed using a 0.1 threshold. Since these are static images, the number of consecutive frames is not needed. The structure of the right panel of Figure 2 was followed when obtaining the average EAR values, and Figure 4 was used to obtain the 13 dimension feature vector for each average EAR value. Then, the model was trained on these feature vectors, where it learned to classify the images as open or closed. To get the optimal hyperparameters for the model, we used the GridSearch Method with a 5-fold cross validation [19].

To evaluate the model, each frame from the eye-stream video contains the average EAR value feature vector with 13 values (using Figure 2 and Figure 4), and the model predicts whether the frame is open or closed using the feature vector as input. Thus, a blink will have occurred if there are successive frames with closed classifications.

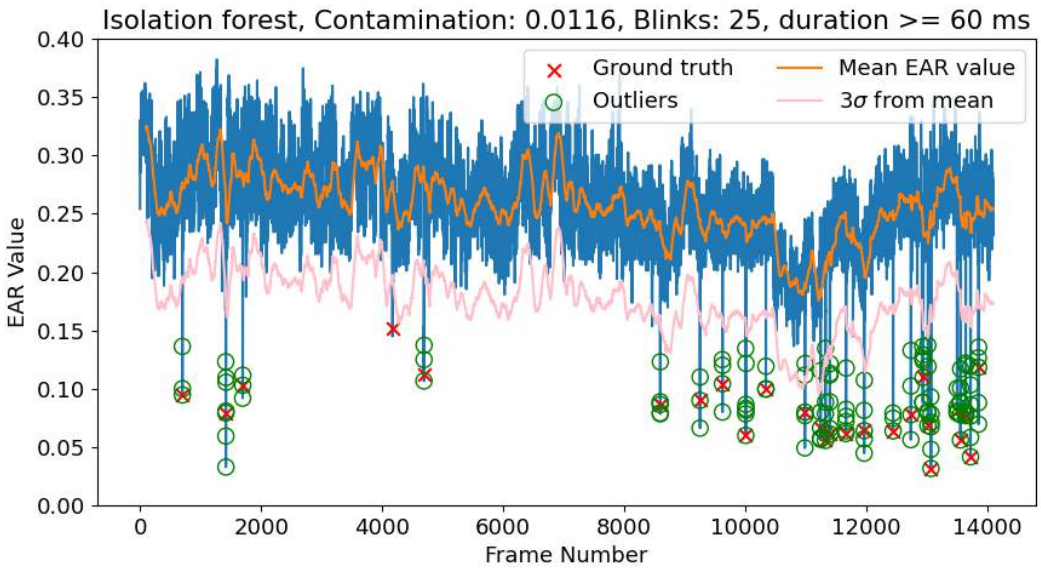
2.4. Unsupervised Learning Method (Isolation Forest)

The noisy nature of the EAR values makes it difficult to establish a threshold value to differentiate what is and is not a blink. Indeed, the EAR values in turn may be sensitive to the individual, motivating a departure from a static threshold. An unsupervised method is also presented as a means to establish a threshold for when a blink occurs.

This approach is based on interpreting blinks as an anomalous occurrence, since eyes are typically open while awake. This interpretation encourages the application of outlier detection algorithms to annotate blinks.

The particular technique chosen is the Isolation Forest algorithm, which is based on an ensemble of decision trees, where anomalies are said to have a *shorter path length* than normal points. The path length is determined by selecting a random data record and isolating into a partition. If the data point is “close" to other data points, then more partitioning, or a longer path length may be required. Conversely, isolated data points

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2022 doi:10.20944/preprints202203.0200.v1



6.png

Figure 5. Isolation forest applied onto EAR values to determine 'outliers' (i.e. blinks) . *Contamination* is a pre-determined proportion of outliers. A blink is classified as such if at least three consecutive frames (time resolution of data is 20 ms) were determined to be outliers.

require fewer partitions, thus a shorter path length [20]. Furthermore, this method is well suited for multi-modal datasets and has low computational overhead.

Although the method does not require a labelled training dataset, it contains a parameter which the algorithm is quite sensitive to: contamination, i.e. the proportion of outliers in the dataset. To overcome a similar issue as with a static threshold for EAR values, contamination has the following automatic estimation. A simple moving average (SMA) of window size of a 100 frames is established. A first order contamination ratio is calculated as follows: the number of data points three standard deviations (3σ) away from the SMA mean to the total number of frames. This approach will enable the contamination value to adapt to the data as opposed to a static threshold. Results can be seen in Figure 6.

To avoid misclassifying sudden troughs in data as blinks, we require also that at least three consecutive frames are classified as outliers. The average duration of a blink is 100-400 ms, and given that the frame rate of the eye-stream is 50 fps, this methodology will be sensitive to eye movement longer than 60 ms.

3. Results

In this section, we evaluate the performance of the given three methods to blink detection. Codes for implementing the techniques were written in Python and are available at the GitHub repository [21].

To assess classifier performances, we applied the three methods to the same eye-stream video. The video contains a grayscale recording of a participant's eyes at two different angles for each eye, while naturally scrolling through their personal Twitter feed.

Blink ground truths are obtained manually from direct observation of each from on the eye-stream video. A blink was defined as a singular frame, i.e. the frame at which the eyes are *completely* closed. This frame was typically preceded by three to four frames in which the eyes were rapidly closing, and proceeded by a longer duration of opening the eyes. There were also a few instances where the eyes were closing but did not completely close, which were not labeled as blinks.

The classification methods are evaluated via True Negative (TN), False Negative (FN), False Positive (FP), and True Positive (TP) counts based on pairwise comparisons with the ground truth. Furthermore, the occurrence of a blink within a $\pm$ 6 frame is accepted. For

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 March 2022 doi:10.20944/preprints202203.0200.v1

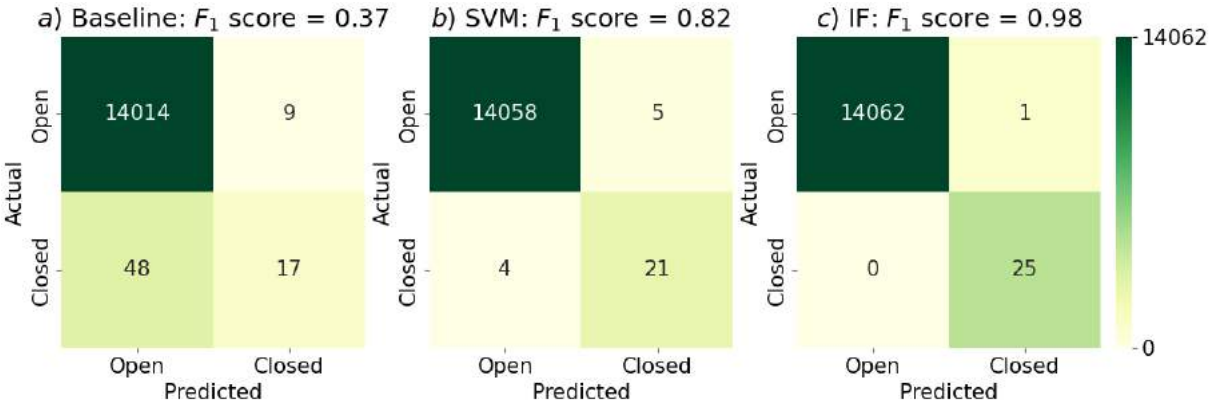


Figure 6. Confusion matrices of the results: **a)** *Baseline* method with a static EAR value threshold, **b)** Supervised prediction with *SVM*, **c)** Unsupervised outlier detection with *IF*. Prediction was accepted as a blink if within $\pm$ 6 frames from a ground truth frame. 26 blinks were observed manually.

example, if a ground truth blink is labeled at frame 700, any true blink flag between 693 - 707 will be recorded as a TP.

An F_1 score, the harmonic mean of precision and sensitivity, is also calculated for each method. Defined as:

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}, \tag{2}$$

It is a measure of relative performance, and is not skewed by the large number of TNs.

As evidenced from the baseline method, a static threshold on EAR values is insufficient to predict blinks accurately. The SVM method performs markedly better, with an F_1 score is 0.82. It should also be noted that due to the rolling nature of the SVM method, it makes 12 less predictions ($\pm$ 6) than other methods. IF method has performed best, with only one misclassification.

4. Discussion

EAR values provide an excellent pathway to detect blinks. However, they can be noisy and require more processing to accurately identify when a blink occurs. This also requires a consistent definition of what a blink is. For the purposes of this study, a blink was defined to be a singular instance of time, when the eyes were *completely* closed. Initially, this required a tedious manual labeling of the frames but through this study, we hope to automate the process of blink detection to a high degree of accuracy.

A static threshold on noisy EAR values (as demonstrated on the Baseline method) misclassified numerous non-blink frames as blinks. SVM method fared much better, but IF performed at the highest degree of accuracy and precision. Furthermore, the IF method does not require a labeled, ground-truth dataset.

This study encourages further analysis into types and phases of blinks and whether they can be reliably classified based on an observed blink duration. As striatal dopamine levels are related spontaneous blink rate, one may allocate a cognitive activation level based on blink frequency. Additionally, accurate blink detection will facilitate removal of blinking artifacts present in EEG data.

As a final note, although these algorithms may apply to other eye-tracking systems, only data from the Tobii Pro Glasses 2 is evaluated here (i.e. the eye-stream video). Therefore, the efficacy of these algorithms with other data requires further investigation. It is our hope that this work can be used to further our understanding of blinking and its association with cognition.

5. Conclusions

Blinks are a powerful, easily accessible, non-invasive way to identify information such as dopaminergic activity and stress indicators, thus it is important to establish an automated detection paradigm. In this paper, we proposed an unsupervised learning method built on top of an already existing technology to make detection more robust. The *Isolation Forest* method is an intuitive, lightweight approach that a) has shown to be more precise and accurate than other methods and b) is unsupervised, thus eliminating the need for a labeled dataset.

Author Contributions: Conceptualization, B.F., A.S., and S.T.; methodology, B.F., A.S., and S.T.; software, B.F. and A.S.; validation, B.F., A.S., and S.T.; formal analysis, B.F., A.S., and S.T.; investigation, B.F., A.S., and S.T.; resources, D.J.L.; data curation,B.F., A.S., and S.T; writing—original draft preparation, B.F., A.S., and S.T; writing—review and editing, B.F., A.S., S.T, J.W., and D.J.L.; visualization, B.F., A.S., and S.T.; supervision, D.J.L.; project administration, D. J. L.; funding acquisition, D. J. L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the following grants: 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. Support from the University of Texas at Dallas Office of Sponsored Programs, Dean of Natural Sciences and Mathematics, and Chair of the Physics Department are gratefully acknowledged. The authors acknowledge the OIT-Cyberinfrastructure Research Computing group at the University of Texas at Dallas and the TRECIS CC* Cyberteam (NSF 2019135) for providing HPC resources that contributed to this research (https://utdallas.edu/oit/departments/circ/, accessed 02/22/2022.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data used https://zenodo.org/record/5874701#.YhPxrJPMJb8

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

EEG	Electroencephalography
EAR	Eye Aspect Ratio
SVM	Support Vector Machine
IF	Isolation Forest

References

1. van der Wel, P.; van Steenbergen, H. Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin and Review* **2018**, *25*, 2005–2015. doi:10.3758/s13423-018-1432-y.
2. Bradley, M.M.; Miccoli, L.; Escrig, M.A.; Lang, P.J. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology* **2008**, *45*, 602–607. doi:10.1111/j.1469-8986.2008.00654.x.
3. Kahnemann, D.; Beatty, J. Pupillary responses in a pitch-discrimination task. *Perception Psychophysics* **1967**, *2*, 101–105. doi:10.3758/BF03210302.
4. Beatty, J.; Kahneman, D. Pupillary changes in two memory tasks. *Psychonomic Science* **1966**, *5*, 371–372. doi:10.3758/BF03328444.
5. Soukupová, T.; Cech, J. Eye-Blink Detection Using Facial Landmarks. *Proceedings of the 21st Computer Vision Winter Workshop* **2016**, pp. 22–29.
6. Sakai, T.; Tamaki, H.; Ota, Y.; Egusa, R.; Inagaki, S.; Kusunoki, F.; Sugimoto, M.; Mizoguch, H. Eda-based estimation of visual attention by observation of eye blink frequency. *International Journal on Smart Sensing and Intelligent Systems* **2017**, *10*, 296–307. doi:10.21307/ijssis-2017-212.

7. Chen, S.; Epps, J. Using task-induced pupil diameter and blink rate to infer cognitive load. *Human-Computer Interaction* **2014**, *29*, 390–413. doi:10.1080/07370024.2014.892428.
8. Agarwal, M.; Sivakumar, R. Blink: A Fully Automated Unsupervised Algorithm for Eye-Blink Detection in EEG Signals. *2019 57th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2019* **2019**, pp. 1113–1121. doi:10.1109/ALLERTON.2019.8919795.
9. Espinosa, J.; Pérez, J.; Mas, D. Comparative analysis of spontaneous blinking and the corneal reflex. *Royal Society Open Science* **2020**, *7*. doi:10.1098/rsos.201016rsos201016.
10. Jongkees, B.J.; Colzato, L.S. Spontaneous eye blink rate as predictor of dopamine-related cognitive function—A review. *Neuroscience and Biobehavioral Reviews* **2016**, *71*, 58–82. doi:10.1016/j.neubiorev.2016.08.020.
11. Groman, S.M.; James, A.S.; Seu, E.; Tran, S.; Clark, T.A.; Harpster, S.N.; Crawford, M.; Burtner, J.L.; Feiler, K.; Roth, R.H.; Elsworth, J.D.; London, E.D.; Jentsch, J.D. In the blink of an eye: Relating positive-feedback sensitivity to striatal dopamine d2-like receptors through blink rate. *Journal of Neuroscience* **2014**, *34*, 14443–14454. doi:10.1523/JNEUROSCI.3037-14.2014.
12. Nieoullon, A. Dopamine and the regulation of cognition and attention. *Progress in Neurobiology* **2002**, *67*, 53–83. doi:10.1016/S0301-0082(02)00011-4.
13. Wise, R.A. Dopamine, learning and motivation. *Nature Reviews Neuroscience* **2004**, *5*, 483–494. doi:10.1038/nrn1406.
14. Goto, Y.; Grace, A.A. Dopaminergic modulation of limbic and cortical drive of nucleus accumbens in goal-directed behavior. *Nature Neuroscience* **2005**, *8*, 805–812. doi:10.1038/nn1471.
15. Terhune, D.B.; Sullivan, J.G.; Simola, J.M. Time dilates after spontaneous blinking. *Current Biology* **2016**, *26*, R459–R460. doi:10.1016/j.cub.2016.04.010.
16. Cech, J.; Soukupova, T. Real-Time Eye Blink Detection using Facial Landmarks. *Center for Machine Perception, Department of Cybernetics Faculty of Electrical Engineering, Czech Technical University in Prague* **2016**, pp. 1 – 8, [arXiv:1011.1669v3].
17. Kazemi, V.; Sullivan, J. One millisecond face alignment with an ensemble of regression trees. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2014**, pp. 1867–1874. doi:10.1109/CVPR.2014.241.
18. Noble, W.S. What is a support vector machine? *Nature Biotechnology* **2006**, *24*, 1565–1567. doi:10.1038/nbt1206-1565.
19. Syarif, I.; Prugel-Bennett, A.; Wills, G. SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance. *TELKOMNIKA (Telecommunication Computing Electronics and Control)* **2016**, *14*, 1502. doi:10.12928/telkomnika.v14i4.3956.
20. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation Forest **2008**. pp. 413–422. doi:10.1109/ICDM.2008.17.
21. Arjun Sridhar.; Shawhin Talebi.; Ashen Fernando. mi3nts/tobiiBlinkDetection: Blink detection algorithms for Tobii Pro Glasses 2 data.

 doi:10.20944/preprints202203.0200.v1

Decoding Physical and Cognitive Impacts of PM Concentrations at Ultra-fine Scales

Shawhin Talebi (shawhintalebi@gmail.com)
The University of Texas at Dallas https://orcid.org/0000-0002-9841-6703
David Lary
The University of Texas at Dallas
Lakitha Wijeratne
The University of Texas at Dallas
Bharana Fernando
The University of Texas at Dallas
Tatiana Lary
The University of Texas at Dallas
Matthew Lary
The University of Texas at Dallas
John Sadler
The University of Texas at Dallas
Arjun Sridhar
The University of Texas at Dallas
John Waczak
The University of Texas at Dallas
Adam Aker
The University of Texas at Dallas
Yichao Zhang
The University of Texas at Dallas

Physical Sciences - Article

Keywords: Biometrics, Particulate Matter, Machine Learning

Posted Date: March 29th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1499191/v1

Decoding Physical and Cognitive Impacts of Particulate Matter Concentrations at Ultra-fine Scales

Shawhin Talebi[1*], David J. Lary[1], Lakitha O. H. Wijeratne[1†], Bharana Fernando[1†], Tatiana Lary[1†], Matthew D. Lary[1†], John Sadler[1†], Arjun Sridhar[1††], John Waczak[1††], Adam Aker[1††] and Yichao Zhang[1††]

[1*]Hanson Center for Space Sciences, The University of Texas at Dallas, 800 W Campbell Rd, Richardson, 75080, TX, USA.

*Corresponding author(s). E-mail(s): shawhin.talebi@utdallas.edu;
Contributing authors: david.lary@utdallas.edu;
†These authors contributed equally to this work.

Abstract

The human body is an incredible and complex sensing system. Environmental factors trigger a wide range of automatic neurophysiological responses. Biometric sensors can capture these responses in real time, providing clues to the underlying biophysical mechanisms. Here we show biometric variables can be used to accurately estimate ultra-local particulate matter concentrations in the ambient environment with high fidelity ($r^2 = 0.91$) and that smaller particles are better estimated than larger ones. Inferring environmental conditions solely from biometric measurements allows us to disentangle key interactions between the environment and the body. A deeper understanding of these interactions can have countless important applications in public health, preventative healthcare, city planning, human performance, and much more. By tapping into our body's 'built-in' sensing abilities, we can gain insights to how our environment influences our physical health and cognitive performance.

Keywords: Biometrics, Particulate Matter, Machine Learning

1 Introduction

Over 4 million premature deaths worldwide were attributed to outdoor air pollution in 2016 [1]. In 2019, 99% of the global population resided in areas that fell short of the World Health Organization (WHO) air quality guidelines [1]. There has been mounting evidence that poor air quality negatively impacts respiratory, cardiovascular, and cerebrovascular health [2–7]. Further, there is emerging evidence on the impact of poor air quality on neurological outcomes including chronic diseases (e.g. Alzheimer's disease and dementia) [2, 8, 9] and acute cognitive impairment [10–14].

Although several large-scale epidemiological studies show the negative effects of air pollution on physical and cognitive health [2–7], these studies largely focused on coarse spatial (∼10 miles) and temporal (∼1 day) scales. Much less research focuses on ultra-local spatial (∼1 m) and temporal (∼10 seconds) scales that make simultaneous environmental and holistic biometric observations of the human physiological responses.

Before an extreme result such as a disease occurs, poor air quality already negatively impacts human physical and cognitive performance [10–14]. Through this work, we investigate how air pollution impacts human performance by examining the relationship between environmental air quality measurements and automatic physiological responses at ultra-fine scales.

This study extends past works that examined interactions of cardiovascular variables such as heart rate (HR), heart rate variability (HRV), and blood pressure (BP) with air quality on fine scales [15–17]. The main contribution of this study is that we augment cardiovascular markers with other biometrics, including electroencephalography (EEG), pupillometry, galvanic skin response (GSR), body temperature, blood oxidation, and respiration rate. This extended set of variables provides insight into both the cardiovascular and cognitive status of the participant. A study of air quality and human physiology at the ultra-local level may shed light on the biophysical mechanisms that underlie their interactions.

2 Results

In this work we used a data-driven experimental paradigm to develop and explore several empirical machine learning models which describe the connection between ambient air particulate matter (PM) concentrations and the biometric variables of an individual breathing that air. Due to logistical constraints imposed by the COVID-19 pandemic, we were only able to collect data from one participant. Additional participants will be included in future research. Two factors, however, mitigate the limited population size in this study. First, the data collection took place over three days, which allowed for contextual variability. Furthermore, the participant repeatedly circled the same trail, allowing for multiple observations of identical spatial positions and 360-degree changes in wind direction angles.

The estimated PM values included: PM_1, $PM_{2.5}$, PM_4, PM_{10}, PM_{Total}, and 45 different PM size bins ranging of 0.18 – 10 μm measured in $\mu g/m^3$, as well as particle count density (dCn) measured in particles per m^3. For model development, 329 biometric predictor variables were available. Two subsets of 9 biometric predictor variables were used in training a set of empirical machine learning models. The first subset includes EEG variables, and the second subset does not. The cognitive effects of air quality can be identified by evaluating predictive models with and without EEG values.

Each machine learning model used was a trained ensemble of decision trees for multi-variate non-linear non-parametric regression with full hyperparameter optimization [18–23]. The empirical models are evaluated using two key metrics. First, the model accuracy assessed using the squared correlation coefficient (r^2) between the model prediction and the true PM values. Second, a ranking of predictor variable importance obtained as the weighted average importance of each predictor across the ensemble.

We first evaluated the six machine learning models for particulate matter (PM) values which estimated: the particle count density (dCn), PM_1, $PM_{2.5}$, PM_4, PM_{10}, and PM_{Total}. 329 biometric predictor variables were used as model inputs including: delta (1 – 3 Hz), theta (4 – 7 Hz), alpha (8 – 12 Hz), beta (13 – 25 Hz), and gamma (25 – 70 Hz) band power densities for each of the 64 EEG electrodes, body temperature, galvanic skin response (GSR), heart rate (HR), heart rate variability (HRV), respiration rate (RR), peripheral capillary blood oxygen saturation (SpO_2), average pupil diameter, the difference between the left and right pupil diameters (anisocoria), and the 3D spatial distance between the left and right pupil centers (vergence eye movement). Then, using an Occam's razor principle, the top 9 important biometric predictor variables were used to train an additional six models for the same PM variables.

The best performing model using the top 9 EEG and non-EEG biometric predictors was for PM_1. This model had the highest accuracy with a validation dataset $r^2 = 0.91$. Comparison plots between estimated and ground truth PM_1 values are given in Figure 1. In the top-left plot, the estimated and true PM_1 concentrations in both the training (blue circles) and validation (red pluses) datasets closely follow to the perfect fit (black) line. In the top-right plot, the quantile-quantile comparison shows the distribution of measured PM_1 values closely resembles the distribution of estimated PM_1 values. Finally, in the bottom plot, the time series of the estimated PM_1 values (dashed red line) tracks very closely to the true values (solid black line) over seven different trials spanning three separate days.

The performance of the PM_1 and five other PM models in this cohort are ranked in the left panel of Figure 2. The training and independent validation dataset performances are plotted in blue and orange, respectively, and sorted in descending order of independent validation performance. As previously discussed, PM_1 measured in $\mu g/m^3$ was best reproduced by the 9 biometric predictors (validation $r^2 = 0.91$). The empirical models based on

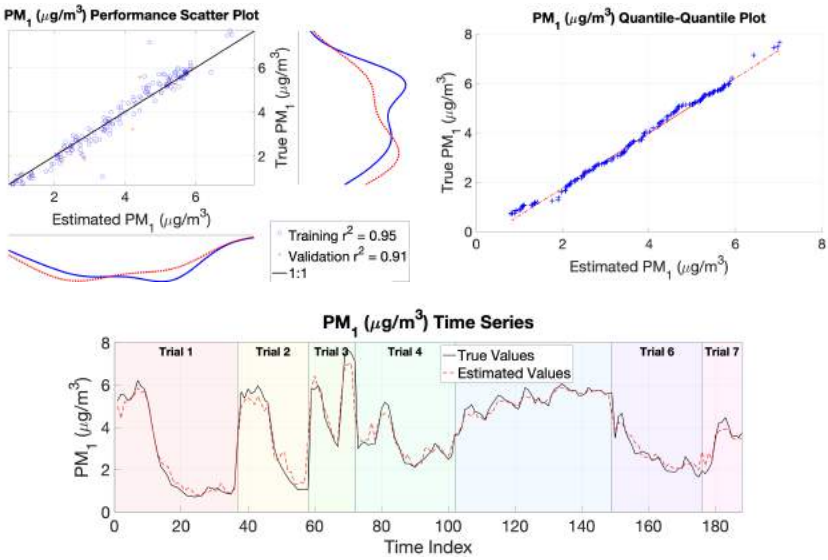


Fig. 1 Top performing model (PM_1) plots comparing predict and ground truth values. (**Top − Left**) scatter plot of true versus predicted PM_1 values. A perfect fit is indicated by the 1:1 line shown in black. Training data are shown as blue circles and validation data are plotted as red pluses. (**Top − Right**) quantile-quantile plot of true versus predict PM_1 values. Identical true and predicted distributions would results in a perfect y=x line. (**Bottom**) Time series plot of true PM_1 values (solid black line) and predicted PM_1 values (dashed red line).

the same biometric predictors were less able to accurately estimate the larger PM_{10} (validation $r^2 = 0.67$) values and PM_{Total} (validation $r^2 = 0.72$) which is dominated by PM_{10} due to the larger masses. The poor performance of these models could be explained the fact that there are significantly fewer large particles than small particles, and thus the larger particles are not as well mixed as the far more numerous and well mixed smaller particles. Because of their greater bulk, larger particles settle more quickly. As a result, the concentrations of large particles collected by the survey vehicle and those inhaled by the subject a few meters away are likely to differ more than for the smaller particles. Second, it's possible that the larger particles have less of an impact on the participant's physical and cognitive state because they are less likely to penetrate deeply into the respiratory and circulatory systems [26].

Each of the six empirical machine learning models has an associated predictor importance ranking, which quantifies the role of individual input predictor variables in estimating the respective PM target variable. The aggregated ranking of top predictors, shown in the right plot in Figure 2, elucidates which biometric variables are most helpful to the empirical models in discerning PM values. The most important predictor variable in estimating PM values was the body temperature measured at the participant's right temple. Surprisingly, the respiratory variable HRV played less of a role. Other important biometrics included GSR and the distance between the pupil centers of the eyes. GSR is

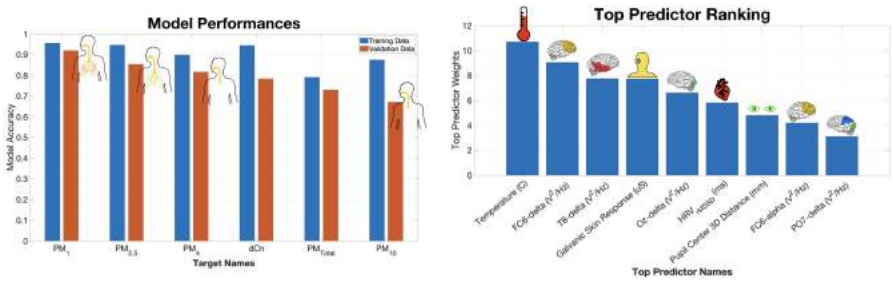


Fig. 2 Summary of empirical PM concentration models estimated from 9 EEG and non-EEG biometric predictor variables. (**Left**) Ranking of model performance defined as squared correlation coefficient between predicted and true PM values. Training and validation dataset performances for each model are shown in blue and orange, respectively. Sorting is based on validation dataset performance. Overlaid graphics indicate the deposition of the respective PM size bins in the airways [26]. (**Right**) Predictor importance ranking aggregated across all 6 models.

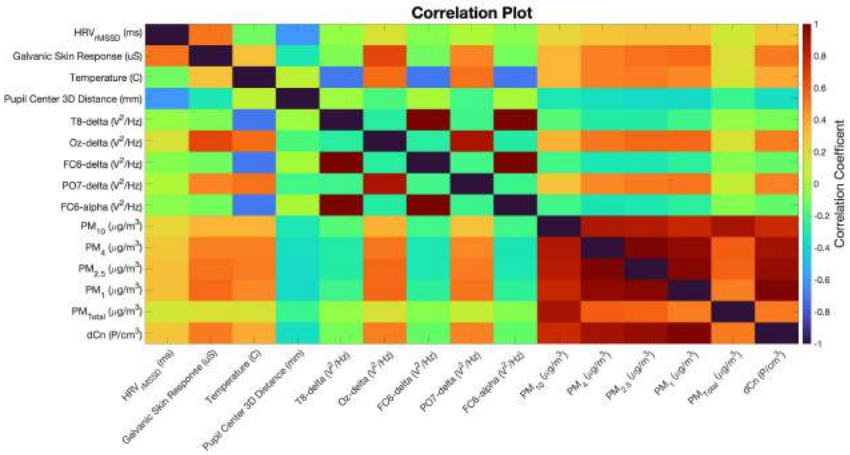


Fig. 3 Correlation plot of top 9 EEG and non-EEG biometric predictor variables, along with 6 target PM variables. Positively correlated variable pairs are indicated by a red box, negatively correlated pairs are shown by blue boxes, and non-correlated pairs have green boxes.

a strong correlate of body temperature. While the distance between the pupil centers is a proxy for vergence eye movements, which have been associated with attentional load and to be a strong predictor of cognitive status [24, 25]. EEG variables found to play an important role in estimating PM values were the delta band (1 – 3 Hz) power densities for the FC6, T8, and Oz electrodes. FC6 is above the frontal cortex on the right side of the head, T8 corresponds to the right temporal lobe, and Oz sits on top of the primary visual cortex.

Correlations between predictor and target variables are visualized as a color filled correlation plot in Figure 3. As seen by the red-orange streaks in the bottom-left and top-right of the correlation plot, HRV, GSR, temperature, and the delta power density of the Oz and PO7 electrode signals have strong

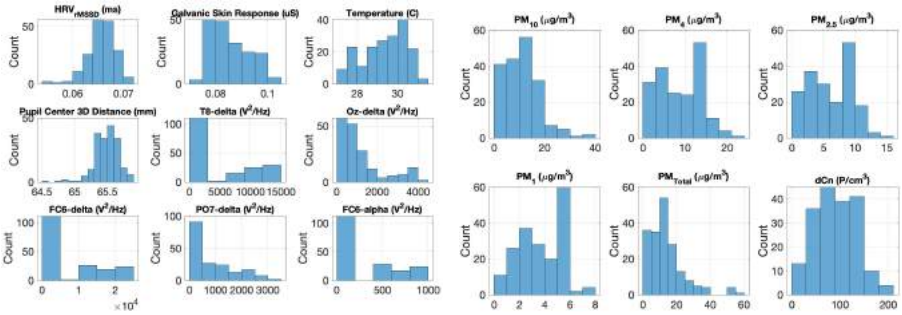


Fig. 4 (**Left**) Histograms of 9 EEG and non-EEG predictor variables. Plots are titled by variable name and its physical units. (**Right**) Histograms of 6 different PM target variables variables. Plots are titled by variable name and its physical units.

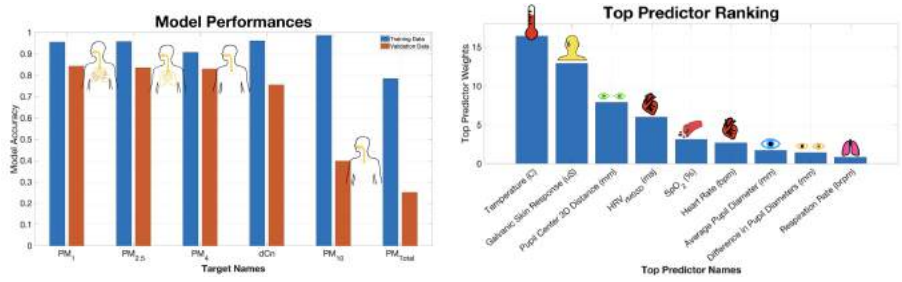


Fig. 5 Summary of empirical PM concentration models estimated from 9 non-EEG biometric predictor variables including eye tracking, respiratory, and other physiological variables. (**Left**) Ranking of model performance defined as squared correlation coefficient between predicted and true PM values. Training and validation dataset performances for each model are shown in blue and orange, respectively. Sorting is based on validation dataset performance. Overlaid graphics indicate the deposition of the respective PM size bins in the airways [26]. (**Right**) Predictor importance ranking aggregated across all 6 models.

positive correlations with all target variables except PM_{Total}. In other words, as these predictor variables increase, so do the corresponding PM target variables. PM target variables show the greatest negative correlation with the 3D spatial distance between left and right pupil centers. Suggesting that the pupils tend to converge with an increase in PM concentrations. Lastly, of all the target variables, PM_{Total} is most strongly correlated with PM_{10} values, which reflects the strong contribution of PM_{10} particles to PM_{Total}.

Histograms for both predictor and target variables are displayed in Figure 4. Plots are titled by the variable name and its respective physical units. From the target PM variable histograms in the right plot of Figure 4, the mass scales of different particle sizes are evident. Namely, the larger sized PM_{10} particles vary over a much larger range ($0 - 40\ \mu g/m^3$) than the smaller PM_4 ($0 - 20\ \mu g/m^3$), $PM_{2.5}$ ($0 - 15\ \mu g/m^3$), and PM_1 particles ($0 - 8\ \mu g/m^3$). This further explains the strong influence of PM_{10} values on PM_{Total}.

Next, an additional set of six empirical machine learning models for the same set of PM targets (dCn, PM_1, $PM_{2.5}$, PM_4, PM_{10}, and PM_{Total}) were

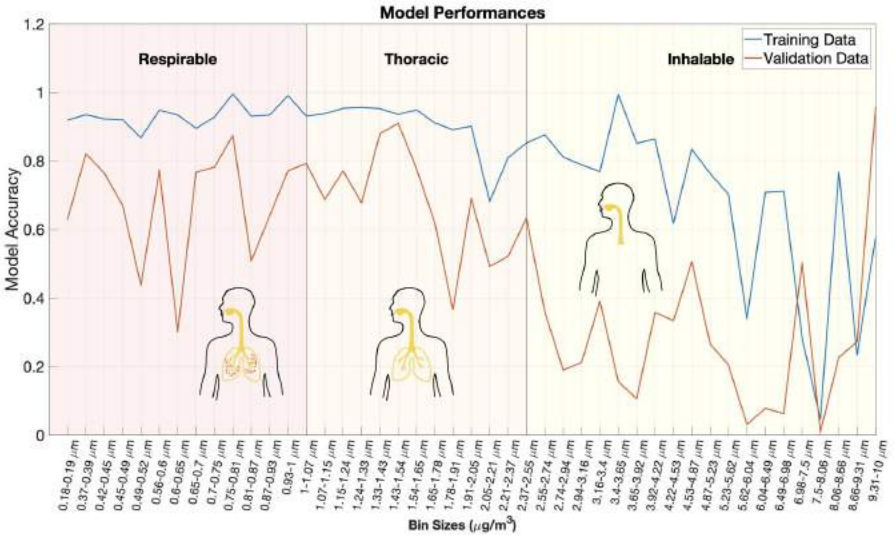


Fig. 6 Model accuracies plotted against bin size. 45 separate PM models were trained for size bins ranging from 0.18 to 10 micrometers. PM values were estimated solely from 9 non-EEG biometric variables. Training dataset performance is plotted as a blue line and validation dataset performance is plotted in orange. A clear drop in model performance is observed between 2 – 3 micrometers. Overlaid graphics indicate the deposition of the respective PM size bins in the airways [26, 27].

evaluated, except this time the PM targets were estimated from 9 non-EEG biometric predictor variables (body temperature, GSR, HR, HRV, RR, SpO_2, average pupil diameter, difference between left and right pupil diameters, and the 3D spatial distance between left and right pupil centers).

The model performance ranking for the six empirical PM models estimated from the 9 non-EEG biometric predictor variables is shown in the left panel of Figure 5. We see that the smaller particles are better estimated by the non-EEG biometrics. Again, this result may be due to better mixing of smaller particles or to deeper penetration of those particles into the respiratory system or both.

Comparing the performance rankings in Figure 2 and Figure 5, there are clear changes in model accuracies. All models with the exception of PM_4 exhibit a drop in performance. The largest drop occurs for the already poor performing PM_{Total} (drop in validation $r^2 = 0.47$) and PM_{10} (drop in validation $r^2 = 0.28$) models.

There is overlap between the importance rankings of Figure 2 and Figure 5. In both cases, body temperature is the most significant predictor of the PM values. Additionally, GSR maintains its order in the ranking as the 2nd most important non-EEG predictors. Although respiratory variables such as HRV and HR appear in the top six of the importance ranking, these variables trail behind temperature, GSR, and the distance between the eye pupil centers.

Fig. 7 Data collection images. (**Left**) Custom made backpack to house biometric devices and recording computer. (**Middle**) Participant and environmental survey vehicle riding in tandem during data collection. (**Right**) Environmental sensors organized in trunk of electric survey vehicle.

The observation that smaller particles are better estimated than larger sized particles, is explored further by evaluating model performances for finer scaled size bins. Here, 45 models were trained to estimate different PM size bins ranging from 0.18 to 10 micrometers using the 9 non-EEG biometrics listed above. Model accuracy is plotted against bin size in Figure 6. Training and validation accuracies are plotted as blue and orange lines, respectively. The regional depositions of each particle size bin is indicated by a label and background shading [26, 27]. The smallest particles (PM_1) are classified as respirable and can penetrate to the alveoli. The next smallest size bin is thoracic ($PM_{2.5}$) which consists of particle penetrating into the bronchioles. The largest size bin are the inhalable particles (PM_{10}) which can enter into the nose, mouth, and trachea.

There is a clear drop in both training and validation dataset accuracies for size bins between 2 to 3 micrometers, corresponding to thoracic and inhalable particles. For particle size bins above this drop, there is large degree of variation in model performances, however most have poor performance with a validation r^2 below 0.4. While the results may imply that smaller particles have a greater impact on physiological systems due to their deeper deposition, that conclusion cannot be reached based upon the present data. The drop in performance for larger particles may be explained in part or completely by the fact that smaller particles are more plentiful and better mixed. An evaluation of the relative contributions of each of these factors requires further investigation.

3 Materials & Methods

3.1 Holistic Sensing

The data in this study are a subset of a holistic biometric and environmental sensing paradigm. The aim of holistic sensing is to capture all relevant information about a system of interest. The full sensor array includes biometric monitors such as: electroencephalography (EEG), eye tracking glasses, electrocardiography (ECG), galvanic skin response (GSR), body temperature, blood oxygen saturation, and heart rate (Figure 8), in addition to environmental factors such as: particulate matter, chemical composition of air, temperature,

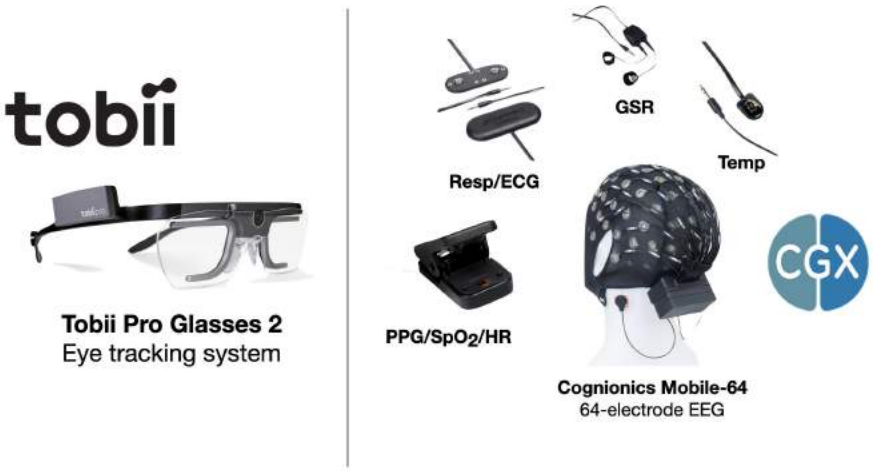


Fig. 8 Biometric sensing systems. (**Left**) Tobii Pro Glasses 2 eye tracking system. This instrument performs eye tracking data, pupillometry, and provides two videos streams of the participant's POV and eyes, respectively. (**Right**) Cognionics Mobile-64 and AIM2 systems. Sensing suite includes 64-electrode EEG, PPG which measures SpO_2 and HR, respiration/ECG sensors, GSR, and temperature probe.

pressure, humidity, visible light spectrum, and more (Figure 9). After processing raw sensor recordings, the full sensor array has a feature space approaching 20,000 variables (∼16,500 biometric and ∼2,000 environmental). In the present study we focus on a relatively small subset, consisting of 329 biometric and 51 environmental variables.

The biometric sensing suite used in this research aims to comprehensively capture the physiological and cognitive status of the participant, without restricting the participant's actions, movements, or decision making. The goal is to gather the maximum amount of information with minimal interruption of normal behaviors. Biometric sensors are placed on the participant in such a way to allow for unrestricted mobility (Figure 10). Sensor recording units and other devices are organized in a backpack worn by the participant that all together weighs less than 10 lbs (Left panel in Figure 7).

Over 100 biometric markers are measured at sampling rates of 500 Hz and 100 Hz. These quantities are processed to derive over 329 variables for the present analysis. This holistic biometric sensing suite integrates two independent sensing systems (Figure 8). Eye tracking is recorded 100 times a second using the Tobii Pro Glasses 2. Data from the glasses produced average pupil diameter, the difference in pupil diameter between left and right eyes, and the 3D spatial distance between pupil centers. All other biometric data are measured 500 times a second using the Cognionics Mobile-64 and AIM2 systems. These systems include a 64-electrode EEG, temperature sensor, respiration sensor, Photoplethysmogram (PPG), and galvanic skin response (GSR) measurement. Heart rate and SpO_2 values are automatically computed by the

Fig. 9 Images of environmental sensing systems. Fidas® Frog Fine Dust Monitoring System measures particulate matter concentrations at 100 different size bins. The AIRMAR 220WX WeatherStation® Instrument samples barometric pressure, wind speed and direction, ambient temperature, and more. The 2B Technologies Black Carbon Photometer measures atmospheric black carbon particulates using long-path photometry. The 2B Technologies Model 205 Dual Beam Ozone sensor is a UV-based ozone monitor. The Konica Minolta CL-500A Illuminance Spectrometer measures the spectral irradiance from 360 to 780 nm at every nanometer. The portable mass spectrometer was constructed by the UNT Laboratory of Imaging Mass Spectrometry and measures charge mass ratios ranging 1 - 300 amu. The 2B Technologies Model 405 nm $NO_2/NO/NO_x$ Monitor™ directly measures atmospheric Nitrogen Dioxide (NO_2) and Nitric Oxide (NO). The LI-COR LI-850 Gas Analyzer measured CO_2 and water vapor in the air.

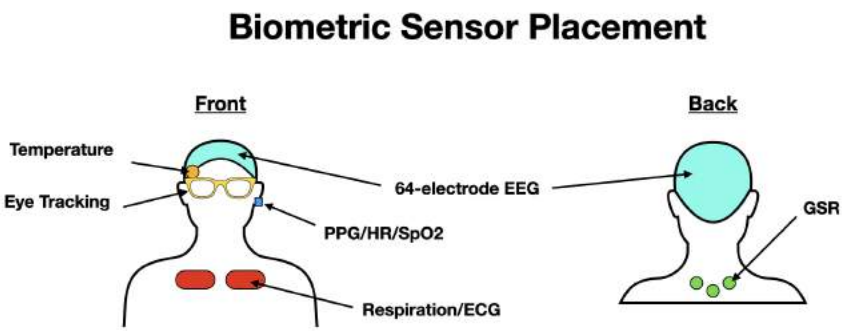


Fig. 10 Schematic of biometric sensor placement on participant. (**Left**) Cartoon of front participant view. The 64-electrode EEG sits on the participant's head. A temperature probe is placed under the EEG cap on the right temple. Eye tracking glasses are carefully placed on participant, avoiding EEG electrodes. PPG sensor is secured to left ear lobe. Respiration sensors are place near the top of the chest. (**Right**) Cartoon of back participant view. GSR sensors are placed below the back of the neck.

AIM2 system using the PPG. Heart Rate Variability (HRV) and Respiration Rate (RR) values are derived from respiration sensor data with a custom

MATLAB script. All biometric data were down-sampled to 1/30 Hz (every 30 seconds) to match particulate matter recordings.

A holistic evaluation of an environmental setting is the ultimate goal of the environmental sensing suite used in these studies. This suite brings together several sensing packages (Figure 9). However, due to its significant societal relevance, for this study we focus on particulate matter (PM) concentrations recorded using the Fidas® Frog fine dust monitoring system. This instrument simultaneously measures PM mass fractions of PM_1, $PM_{2.5}$, PM_4, PM_{10}, and a size distribution within a size range of 0.18 - 100 micrometers as well as the total particle count density (dCn). PM data was recorded at sampling rate of 1 Hz and down-sampled to 1/30 Hz (every 30 seconds).

3.2 Data Collection

Biometric data collection was restricted to a single participant due to logistical constraints arising from the COVID-19 pandemic. However, future works will include data from multiple participants. The small population size in the present study is mitigated by two factors. First, data was collected over three separate days, providing a range of contexts. Additionally, the participant circled the same trail multiple times, offering multiple observations of identical positions and 360-degree changes in wind direction angles.

Data were collected while the participant rode a bicycle in a dynamic outdoor setting. An electric survey vehicle equipped with a suite of environmental sensors followed safely behind the participant during all rides (Middle image in Figure 7. Although several dimensions of the environmental context were sampled (e.g. ambient light, temperature, pressure, mass spectra, etc.), here we focus on the relationship between particulate matter values and biometric variables. Additional relationship will be explored in future works.

Data collection took place in May and June of 2021 at Breckenridge Park located in Richardson, TX over three separate days which included four to five trials per day. The first two trials consisted of two minutes of eyes closed and eye open baseline biometric measurements, respectively. The third trial consisted of a "warm-up" ride, where the participant cycled to a public bike trail in tandem with the electric survey vehicle. Additional trials consisted of the participant repeatedly cycling a one-mile loop on a public bike trail. The participant was free to stop cycling at their discretion. Data collection was halted whenever cycling stopped. If the participant chose to continue, a new data collection trial was initiated.

The complete dataset consists of 188 data records collected every 30 seconds (total time of about 1.5 hours) with 329 biometric predictor variables and 51 PM target variables. Biometric predictor variables include: delta (1 – 3 Hz), theta (4 – 7 Hz), alpha (8 – 12 Hz), beta (13 – 25 Hz), and gamma (25 – 70 Hz) band power densities for each of the 64 EEG electrodes, body temperature, GSR, HR, HRV, RR, SpO_2, average pupil diameter, difference between left and right pupil diameters, and the 3D spatial distance between left and right pupil centers. Environmental PM target variables include: PM_1, $PM_{2.5}$, PM_4, PM_{10},

PM_{Total}, and 45 different PM size bins ranging of 0.18 – 100 μm measured in $\mu g/m^3$, as well as particle count density (dCn) measured in P/cm^3. The data is made publicly available at the Zenodo datastore: https://zenodo.org/record/6326357#.Yieu4RPMJb8.

Ethical approval declarations All experimental protocols were approved by The University of Texas at Dallas Institutional Review Board and informed consent was obtained from the study participant.

3.3 Model Development

All models of PM concentration are obtained by an ensemble of decision trees for regression with a hyperparameter optimization process [18–23]. 90% of the data is used for training, while 10% is help back as an independent validation dataset. Scripts for model training are freely available at the GitHub repository: https://github.com/mi3nts/DUEDARE.

4 Conclusion

The human body and environment form a complex ecosystem. A key aspect of this system is air quality and the effects it has on our bodies. Environmental factors trigger physiological responses that can be detected by holistic biometric sensing. Here we used an ultra-fine holistic sensing paradigm to show particulate matter concentrations in the ambient environment can be accurately estimated using only nine biometric variables. In addition, smaller particles were found to be more accurately estimated. Two potential causes may explain this result. First, smaller particles are much more abundant and well mixed in the ambient environment than larger ones, thus resulting in a greater similarity between particles inhaled by the participant and collected by the survey vehicle. Secondly, smaller particles can deposit into the respiratory system more deeply, and may have a greater impact on the body. Further investigation is needed to assess the relative contributions, if any, of these two factors, since they are not mutually exclusive.

Although the present work shows preliminary findings from a single participant over multiple days, future research will include data from multiple participants. Additionally, several other variables collected (e.g. ambient light, temperature, pressure, mass spectra, etc.) will be evaluated for their physiological interactions. By understanding the key interactions between the environment and the human body, health and performance can be improved across many different domains.

Supplementary information. The data and code has been made publicly available. The full data set is available at the Zenodo data store: https://zenodo.org/record/6326357#.Yieu4RPMJb8 and code is available at the GitHub: https://github.com/mi3nts/DUEDARE.

Declarations

- This research was funded by the following grants: 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. Support from the University of Texas at Dallas Office of Sponsored Programs, Dean of Natural Sciences and Mathematics, and Chair of the Physics Department are gratefully acknowledged. The authors acknowledge the OIT-Cyberinfrastructure Research Computing group at the University of Texas at Dallas and the TRECIS CC* Cyberteam (NSF 2019135) for providing HPC resources that contributed to this research (https://utdallas.edu/oit/departments/circ/, accessed 02/22/2022.
- The authors declare no conflicts of interest
- **Ethical approval declarations** All experimental protocols were approved by The University of Texas at Dallas Institutional Review Board.
- Informed consent was obtained from the study participant.
- Data used in this study is publicly available at the Zenodo datastore: https://zenodo.org/record/6326357#.Yieu4RPMJb8
- Code to reproduce analysis and visualization in this work are available at the GitHub repository: https://github.com/mi3nts/DUEDARE
- Authors' contributions: Methodology, D.J.L., S.T., T.L.; software, S.T.; formal analysis, D.J.L., S.T., T.L., B.F.; data curation, S.T., D.J.L, L.O.H.W., B.F., T.L., M.D.L., J.S., A.S., A.A., Y.Z.; writing—original draft preparation, S.T., D.J.L; writing—review and editing, S.T., D.J.L, B.F., T.L., J.S.; visualization, S.T.; supervision, D.J.L.

References

[1] WHO. Ambient (outdoor) air pollution. Retrieved January 24, 2022, from https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[2] Manisalidis I, Stavropoulou E, Stavropoulos A, Bezirtzoglou E. Environmental and Health Impacts of Air Pollution: A Review. Front Public Health. 2020;8:14. Published 2020 Feb 20. doi:10.3389/fpubh.2020.00014

[3] Orellano, P., Reynoso, J., Quaranta, N., Bardach, A., & Ciapponi, A. (2020). Short-term exposure to particulate matter (PM 10 and PM 2.5), nitrogen dioxide (NO 2), and ozone (O 3) and all-cause and cause-specific mortality: Systematic review and meta-analysis. Environment International, 142. https://doi.org/10.1016/J.ENVINT.2020.105876

[4] Daellenbach, K. R., Uzu, G., Jiang, J., Cassagnes, L. E., Leni, Z., Vlachou, A., Stefenelli, G., Canonaco, F., Weber, S., Segers, A., Kuenen, J. J. P., Schaap, M., Favez, O., Albinet, A., Aksoyoglu, S., Dommen, J., Baltensperger, U., Geiser, M., El Haddad, I., ... Prévôt, A. S. H. (2020). Sources of particulate-matter air pollution and its oxidative potential in Europe. Nature, 587(7834), 414–419. https://doi.org/10.1038/S41586-020-2902-8

[5] Pope, C. A., Brook, R. D., Burnett, R. T., & Dockery, D. W. (2010). How is cardiovascular disease mortality risk affected by duration and intensity of fine particulate matter exposure? An integration of the epidemiologic evidence. Air Quality, Atmosphere & Health 2010 4:1, 4(1), 5–14. https://doi.org/10.1007/S11869-010-0082-7

[6] Brook, R. D., Rajagopalan, S., Pope, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R. V., Mittleman, M. A., Peters, A., Siscovick, D., Smith, S. C., Whitsel, L., & Kaufman, J. D. (2010). Particulate Matter Air Pollution and Cardiovascular Disease. Circulation, 121(21), 2331–2378. https://doi.org/10.1161/CIR.0B013E3181DBECE1

[7] Brook, R. D., Bard, R. L., Kaplan, M. J., Yalavarthi, S., Morishita, M., Dvonch, J. T., Wang, L., Yang, H. Y., Spino, C., Mukherjee, B., Oral, E. A., Sun, Q., Brook, J. R., Harkema, J., & Rajagopalan, S. (2013). The effect of acute exposure to coarse particulate matter air pollution in a rural location on circulating endothelial progenitor cells: results from a randomized controlled study. Inhalation Toxicology, 25(10), 587–592. https://doi.org/10.3109/08958378.2013.814733

[8] Schikowski T, Altuğ H. The role of air pollution in cognitive impairment and decline. Neurochem Int. 2020 Jun;136:104708. doi: 10.1016/j.neuint.2020.104708. Epub 2020 Feb 21. PMID: 32092328.

[9] Kelly FJ, Fussell JC. Air pollution and public health: emerging hazards and improved understanding of risk. Environ Geochem Health. 2015 Aug;37(4):631-49. doi: 10.1007/s10653-015-9720-1. Epub 2015 Jun 4. PMID: 26040976; PMCID: PMC4516868.

[10] Lavy, V., Ebenstein, A., & Roth, S. (2014). The Impact of Short Term Exposure to Ambient Air Pollution on Cognitive Performance and Human Capital Formation. https://doi.org/10.3386/W20648

[11] Chatzidiakou, L., Mumovic, D., & Dockrell, J. (2015). The Effects of Thermal Conditions and Indoor Air Quality on Health, Comfort and Cognitive Performance of Students. https://doi.org/978-0-9930137-3-7

[12] Zhang, X., Chen, X., & Zhang, X. (2018). The impact of exposure to air pollution on cognitive performance. PNAS, 115(37), 9193–9197. https://doi.org/10.1073/pnas.1809474115

[13] Shehab, M. A., & Pope, F. D. (2019). Effects of short-term exposure to particulate matter air pollution on cognitive performance. Scientific Reports, 9(1). https://doi.org/10.1038/s41598-019-44561-0

[14] Künn, S., Palacios, J., & Pestel, N. (2021). Indoor Air Quality and Cognitive Performance. SSRN Electronic Journal. https://doi.org/10.2139/SSRN.3460848

[15] Buteau, S., & Goldberg, M. S. (2016). A structured review of panel studies used to investigate associations between ambient air pollution and heart rate variability. Environmental Research, 148, 207–247. https://doi.org/10.1016/J.ENVRES.2016.03.013

[16] Amoabeng Nti, A. A., Robins, T. G., Mensah, J. A., Dwomoh, D., Kwarteng, L., Takyi, S. A., Acquah, A., Basu, N., Batterman, S., & Fobil, J. N. (2021). Personal exposure to particulate matter and heart rate variability among informal electronic waste workers at Agbogbloshie: a longitudinal study. BMC Public Health 2021 21:1, 21(1), 1–14. https://doi.org/10.1186/S12889-021-12241-2

[17] de Paula Santos, U., Ferreira Braga, A. L., Artigas Giorgi, D. M., Amador Pereira, L. A., Grupi, C. J., Lin, C. A., Bussacos, M. A., Trevisan Zanetta, D. M., Hilário Do Nascimento Saldiva, P., &; Terra Filho, M. (2005). Effects of air pollution on blood pressure and heart rate variability: a panel study of vehicular traffic controllers in the city of São Paulo, Brazil. European Heart Journal, 26 (2), 193–200. https://doi.org/10.1093/EURHEARTJ/EHI035

[18] Breiman, L, Friedman, J H, Olshen, R A, and Stone, C J, 1984, Classification and regression trees: Wadsworth, Inc.

[19] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[20] Tin Kam Ho, The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 832-844, Aug. 1998, doi: 10.1109/34.709601.

[21] Friedman, Jerome H. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, vol. 29, no. 5, Institute of Mathematical Statistics, 2001, pp. 1189–232, http://www.jstor.org/stable/2699986.

[22] Freund, Y. and Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and

System Sciences, 55, 119-139. http://dx.doi.org/10.1006/jcss.1997.1504

[23] Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), e1301. https://doi.org/10.1002/WIDM.1301

[24] Balaban, C. D., Kiderman, A., Szczupak, M., Ashmore, R. C., & Hoffer, M. E. (2018). Patterns of Pupillary Activity During Binocular Disparity Resolution. Frontiers in Neurology, 9(NOV), 990. https://doi.org/10.3389/fneur.2018.00990

[25] Huang, M. X., Li, J., Ngai, G., Leong, H. V., & Bulling, A. (2019). Moment-to-Moment Detection of Internal Thought from Eye Vergence Behaviour. http://arxiv.org/abs/1901.06572

[26] Poh, T. Y., Ali, N. A. T. B. M., Mac Aogáin, M., Kathawala, M. H., Setyawati, M. I., Ng, K. W., & Chotirmall, S. H. (2018). Inhaled nanomaterials and the respiratory microbiome: Clinical, immunological and toxicological perspectives. Particle and Fibre Toxicology, 15(1), 1–16. https://doi.org/10.1186/s12989-018-0282-0

[27] World Health Organization. Occupational and Environmental Health Team. (1999). Hazard prevention and control in the work environment: : airborne dust. World Health Organization. https://apps.who.int/iris/handle/10665/66147