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: October 2021

TYPE OF REPORT: Annual

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

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE October 2021	**2. REPORT TYPE** Annual	**3. DATES COVERED** 30Sep2020-29Sep2021

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** USAMRMC
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	Unclassified	77	**19b. TELEPHONE NUMBER** *(include area code)*

TABLE OF CONTENTS

1. INTRODUCTION:

Our goal is a holistic methodology for providing accurate actionable information on environmental dosimetry for atmospheric composition on fine spatial and temporal scales. The approach uses a state of the art sensor network involving fixed and mobile sensors with real-time cross 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 the active human participants.

2. KEYWORDS:

Dense Urban Environment, Dosimetry, Exposure, Biometrics, Machine Learning

3. ACCOMPLISHMENTS:

What were the major goals of the project?

- Sensor Acquisition & Calibration: 100% Complete
- Electric Vehicle Environmental Sensor Integration: 100% Complete
- Environmental Measurement Campaigns: 90% Complete
- Low-cost sensor calibration and deployment: 100% Complete
- Publication/Conference Presentation. 1 publication appeared, 3 presentations
- Machine Learning Analysis: 25% Complete
- Survey with participants: 30% Complete
- Machine learning analysis linking biometric responses to environmental triggers: 30% Complete

What was accomplished under these goals?

Our environmental surveys of the dense urban environment of the Dallas Fort Worth Metroplex is well underway. We have partnered with local government including Dallas County and the City of Plano.

All publications to date include:

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). 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
5. Yu, Xiaohe, and David J. Lary 2021. "Cloud Detection Using an Ensemble of Pixel-Based Machine Learning Models Incorporating Unsupervised Classification" Remote Sensing 13, no. 16: 3289. https://doi.org/10.3390/rs13163289

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.

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 at https://github.com/mi3nts.

What opportunities for training and professional development has the project provided?

There have been substantial opportunities for training and professional development of the many students involved in this project.

This has included **72 students**: 9 high school students, 53 undergraduate students, and 10 graduate students. They have been involved in sensor construction, sensor calibration using machine learning, research and analysis, writing research papers, presentations at meetings and to community groups, and significant community outreach. The biometric analysis has been the subject of 5 semesters of senior design projects for engineering students, each semester involving a different team of 4-5 students. Likewise the live map displaying the data from sensors cross the DFW area has been the subject of 5 semesters of senior design projects for engineering students, each semester involving a different team of 4-5 students. Our air quality mapping team won first place in the end of semester completion among all the senior design teams. Graduate and undergraduate students were the key senior design team mentors and gained valuable leadership experience in the process.

There has also been an active partnership with the Dallas College community college district, where some of the sensors are located, and Paul Quinn College (the oldest historically black college west of the Mississippi River and the nation's first urban work college) also partnering with sensor deployment.

How were the results disseminated to communities of interest?

1. Extended visit to US SOCOM and SOFWERX in Tampa, FL. Presentations to various SOCOM groups. Appointed United States Special Operations Command Fellow, SOFWERX, J5 Futures Missions Directorate. Awarded a numbered acknowledgement coin.
2. Presentation to General Koeniger Commander of the 711th Human Performance Wing. Awarded an acknowledgement coin.
3. Presentation at the Warrior Human Performance Research Center. Awarded an acknowledgement coin.
4. There is weekly community group interaction, our live map displaying the real-time air quality data (https://www.sharedairdfw.com) is developed with community involvement involving weekly developer meetings, and quarterly partner meetings with the partner community groups, municipalities, counties, school districts and area community colleges involved.
5. Appointed as Adjunct Professor of Military/Emergency Medicine at the Uniformed Services University of the Health Sciences, Bethesda, MD.
6. Documentary of how our sensors led to the closure of an illegal dump, “Disrupt & Dismantle Full Episode: Shingle Mountain” https://www.bet.com/video/disrupt-and-dismantle/season-1/full-episodes/episode-101-shingle-mountain.html

What do you plan to do during the next reporting period to accomplish the goals?

- Conduct more street level surveys of the dense urban environment.
- Conduct joint comprehensive biometric and environmental measurement campaigns with cyclists.
- Complete deployment 24/7 street level sensors.
- Machine learning analysis of data from comprehensive biometric and environmental measurement campaigns with cyclists.

4. **IMPACT:**

What was the impact on the development of the principal discipline(s) of the project?

- We have been told by many people in the Human Performance space that this is the first time that such comprehensive environmental and biometric information has been brought together.
- Based on a literature survey, our calibration study of the pupillary response to light provides the most accurate model to date, and the most comprehensive in terms of wavelength resolution.
- Aspects of this study have led to a follow on robotic sentinel team study for SOFWERX answering the question “is the area safe” that uses the same mass spectrometer on a robotic boat.
- Local government in the Dallas area are now partnering with us thanks to the electric survey car that is part of this project.

What was the impact on other disciplines?

By definition, this project is multidisciplinary.

- The environmental sensing sentinels deployed as part of this project (electric environmental survey car & 24/7 street level sensors) are benefiting local communities in terms of environmental exposure surveys.
- Local law enforcement with the ability to “sniff” meth houses etc.
- Biometric sensing developed in this project is now being used in a SOCOM POTFF project for “live fire” training at Troysgate.
- Two day site visit by several government agencies coordinated by SOFWERX in August 2021.

What was the impact on technology transfer?

Nothing to report so far.

What was the impact on society beyond science and technology?

- Local government in the Dallas area are now partnering with us thanks to the electric survey car that is part of this project for environmental public health protection. This has led to the city of Plano, TX, requesting us to build them a network of 55 street level 24/7 air quality sentinels of the same kind used in this project to deploy across the city of Plano, TX.
- Dallas county is also now partnering with us and linking our live feed data and maps as part of their environmental health protection.
- Several other "preemptive human protection" projects have been spawned thanks to this study. We greatly appreciate your support, thank you, it is making a difference.
- The community group "Downwinders at Risk," the oldest environmental group in Texas, raised funds and have commissioned us to provide a 11 node network (utilizing the same type of 24/7 sensors as in this study) for one of the most polluted communities in south Dallas.
- 72 students have been involved with this project: 9 high school students, 53 undergraduate students, and 10 graduate students. The undergraduate student who's was involved in designing the long range wireless communication for the street level sentinels won an undergraduate research scholar award for this work. The graduate student who did the work on building the pupil dilation models won a Dean's award for his poster on this work.

5. CHANGES/PROBLEMS:

COVID-19 has delayed data collection. Extreme heat and cold experienced this year in Texas has led to some battery issues. There have been some mass spectrometer issues.

Changes in approach and reasons for change

Nothing to Report

Actual or anticipated problems or delays and actions or plans to resolve them

A no-cost extension was granted so that there is more time to complete the participation of human subjects once COVID-19 has passed.

Changes that had a significant impact on expenditures

Nothing to Report

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents

Nothing to Report

Significant changes in use or care of human subjects

Nothing to Report

Significant changes in use or care of vertebrate animals

Not applicable

Significant changes in use of biohazards and/or select agents

Not applicable

6. PRODUCTS:

- **Publications, conference papers, and presentations**

Journal 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). 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
5. Yu, Xiaohe, and David J. Lary 2021. "Cloud Detection Using an Ensemble of Pixel-Based Machine Learning Models Incorporating Unsupervised Classification" Remote Sensing 13, no. 16: 3289. https://doi.org/10.3390/rs13163289

This award was acknowledged. Thank you!

Books or other non-periodical, one-time publications.

Other publications, conference papers and presentations.

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.

- **Website(s) or other Internet site(s)**

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 at https://github.com/mi3nts.

- **Technologies or techniques**

- **Inventions, patent applications, and/or licenses**

- **Other Products**

7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

Shawhin Talebi
Graduate Student
His machine learning research project (part of his graduate study in Physics) led to the publication:

Shawhin Talebi, David J Lary, Lakitha OH Wijerante, Tatiana Lary, Modeling Autonomic Pupillary Responses from External Stimuli using Machine Learning, Biomedical Journal of Scientific & Technical Research, 20 (3), 14,999-15,009, (2019)

Which also won a Dean’s award when it was presented as a poster.

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

Nothing to Report

What other organizations were involved as partners?

Prof. Guido Verbeck's group at the University of North Texas built for us the mass spectrometer. One of only ten of its kind in the world which has performed better than the reference instruments at the Army proving ground trials in 2019. We gratefully acknowledge and appreciate their partnership.

We have partnered with the community group "Downwinders at Risk" and the City of Plano, TX, to provide additional sensors.

We have partnered with the Dallas College Community College District (14 locations) as sensor hosts as well as various school districts and the City of Fort Worth, TX.

8. SPECIAL REPORTING REQUIREMENTS

COLLABORATIVE AWARDS:

QUAD CHARTS:

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

BA170483

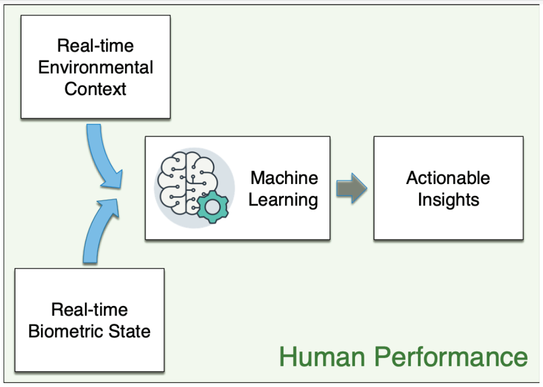
PI: Prof. David J. Lary **Org:** University of Texas at Dallas **Award Amount:** $558,235

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.



Real-time Environmental Context

Real-time Biometric State

Machine Learning

Actionable Insights

Human Performance

Timeline and Cost

Activities CY	2018	2019	2020	2021
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, September, 2021

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

9. APPENDICES:

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 OH Wijerante and Tatiana Lary

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

***Corresponding author's:** Shawhin Talebi, William B. Hanson Center for Space Science, 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 OH Wijerante, 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 ex-ample, changing light intensity provokes a change in the pupil dilation. In the past, formulae for pupil size based on luminance have been de-rived 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 (illuminance) of the wavelengths near 562 nm. This coincides with the peak sensitivity of the longwave photosensitive cones in the retina, which exhibit a maximum absorbance around 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 being missing and the challenge of ending 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.

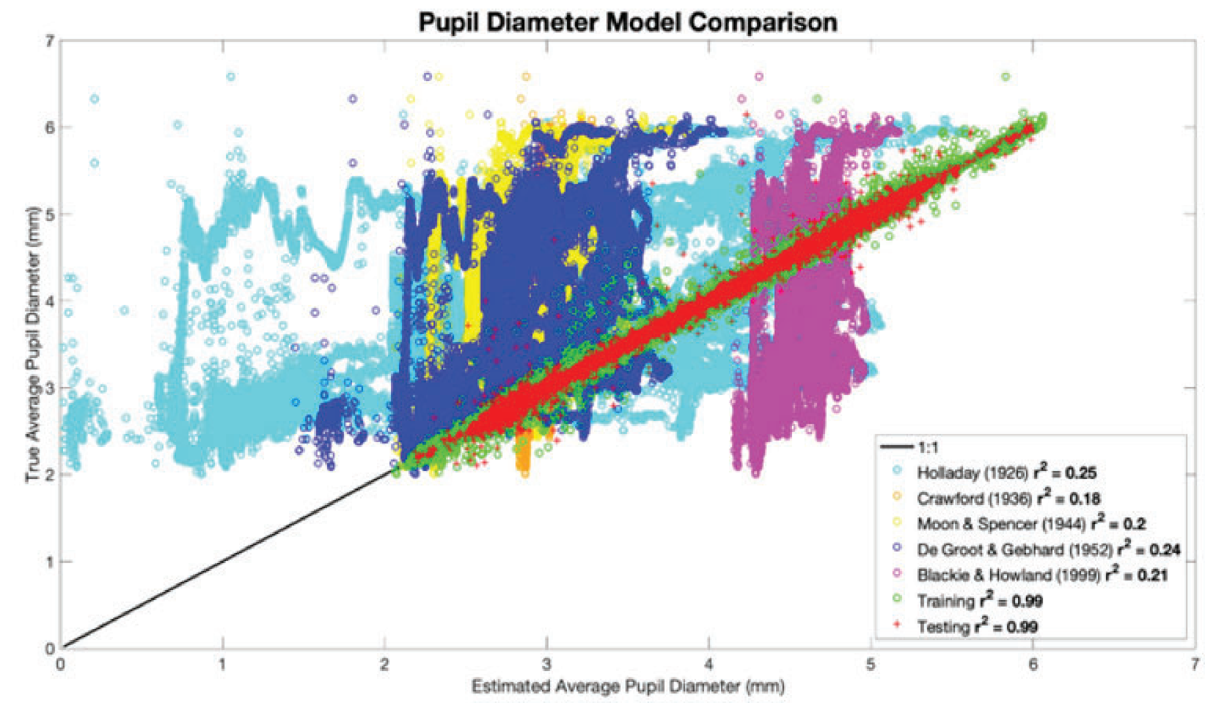


Figure 1: Evaluation and comparison of previous pupil diameter models which utilized a single variable, luminance, showing poor fidelity contrasted with the multivariate empirical machine learning model for the average pupil diameter developed in this study showing good fidelity (foreground green training 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 eject 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 which 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).

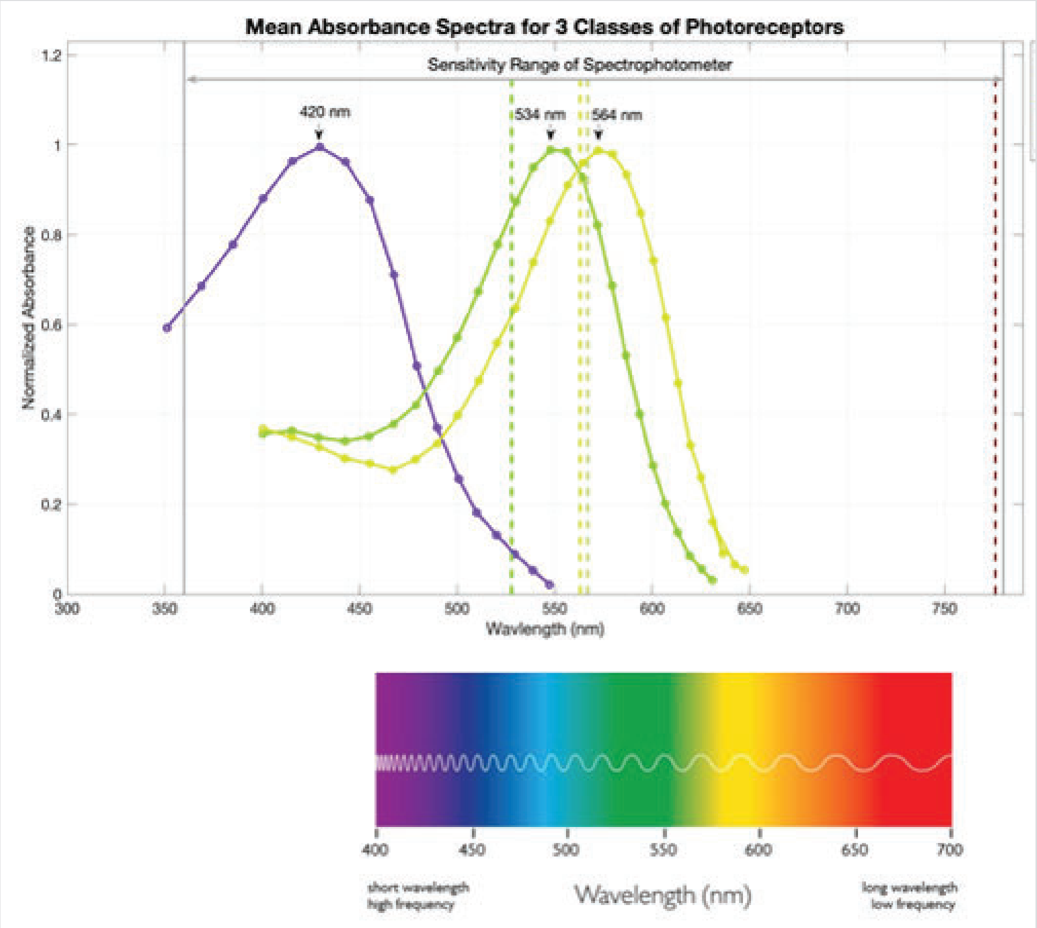


Figure 2: Normalized mean absorbance spectra for long-wave, middle-wave, and short-wave cones. Maximum absorbance values for each class of cones are 420 nm 4.7 nm, 534 nm 3.7 nm, and 564 4.7 nm, respectively. Dashed vertical lines represent the top 4 important predictors taken from the pupil diameter 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 illumi-nance 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 in- dicated 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

b) measurements during late afternoon (≈ 4 PM).

c) Walk 2 measurements dur- ing morning (≈ 8:30 AM) with overcast.

d) 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 (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, r2, 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 predictor Importance 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 3.1 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.

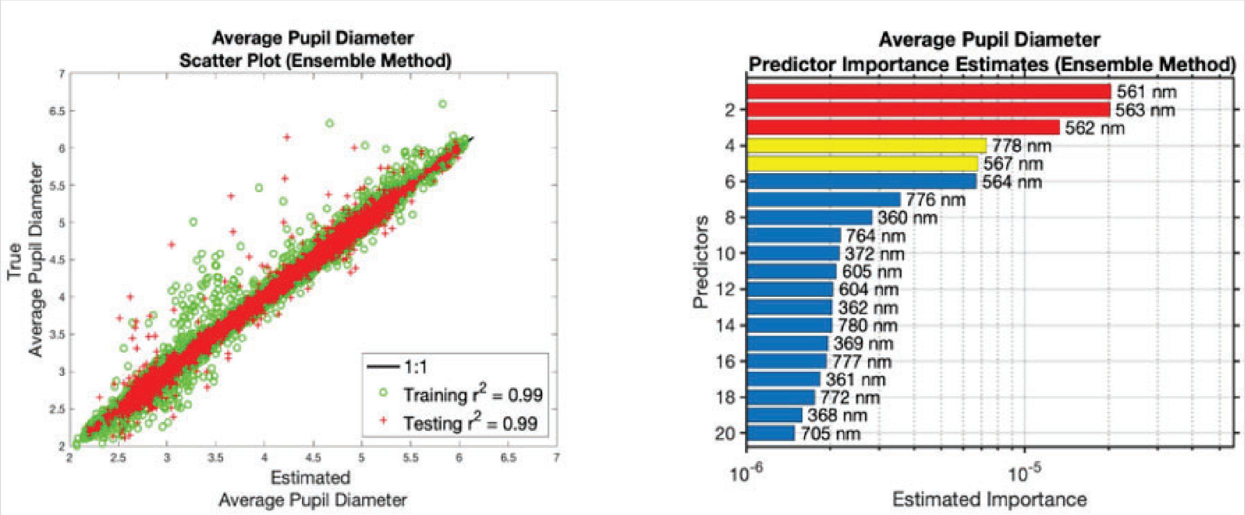


Figure 4: Plots for the average pupil diameter model.

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 irradi-ance 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].

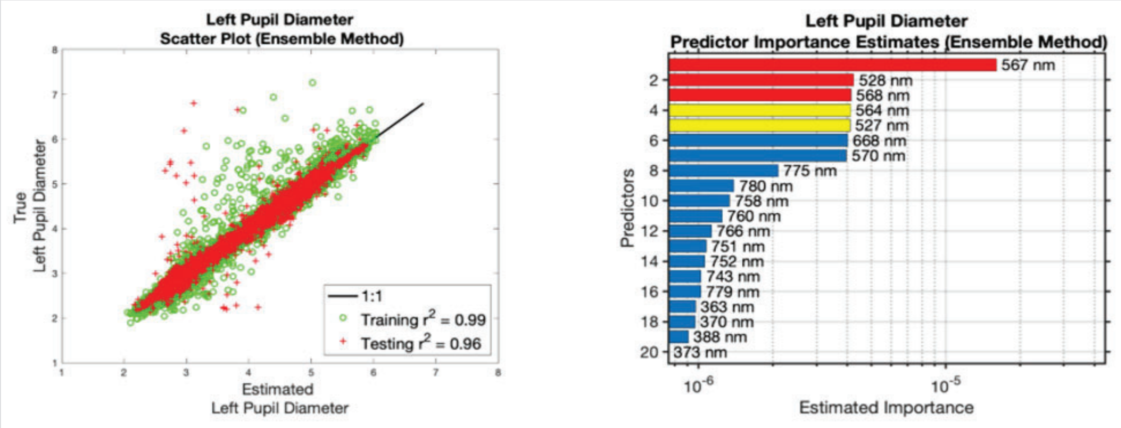


Figure 5: Plots for the left pupil diameter model.

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].

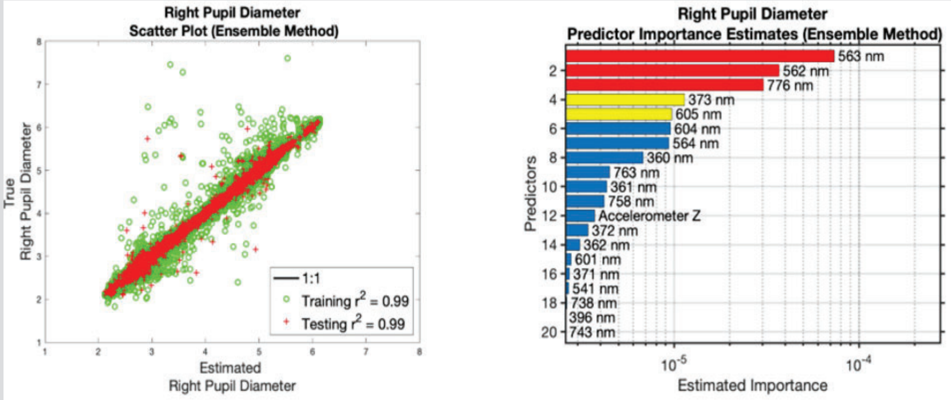


Figure 6: Plots for the right pupil diameter model in millimeters.

a) True vs estimated right pupil diameter.

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 3.4. Clearly the most important predictors for modeling this asymmetry were not available in the training dataset.

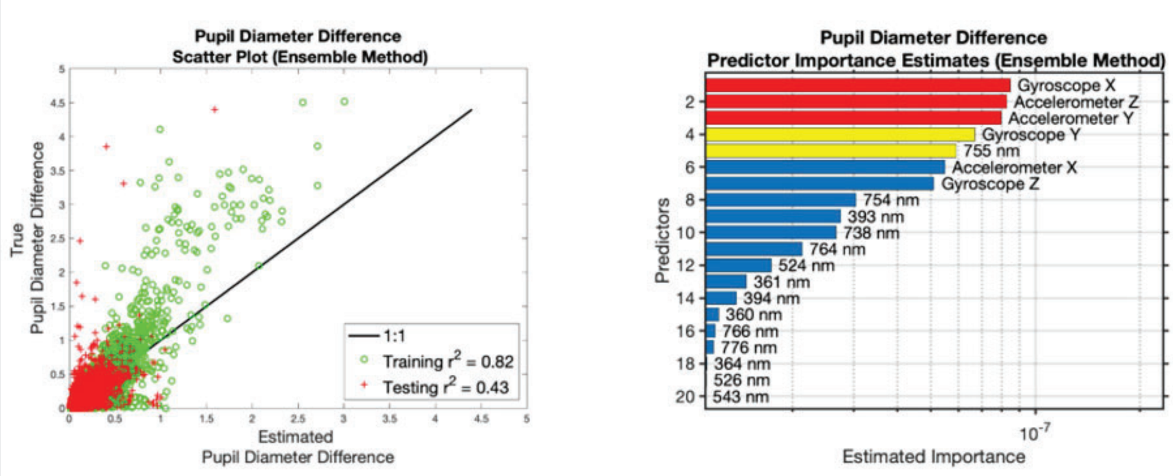


Figure 7: Plots for the pupil diameter difference model.

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].

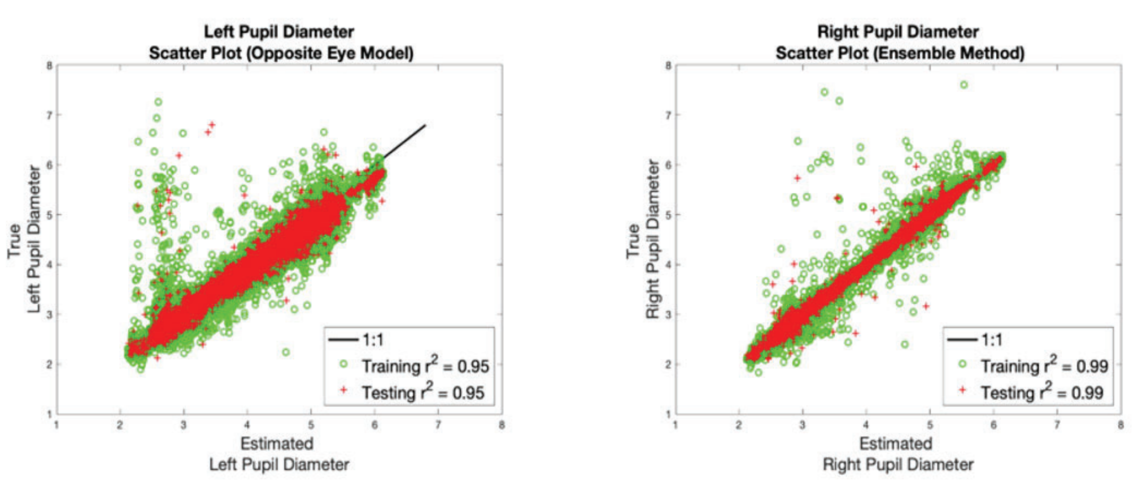


Figure 8: Plots for the pupil diameter prediction using model from opposite eye data. Pupil diameters are in millimeters.

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 some-what 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.

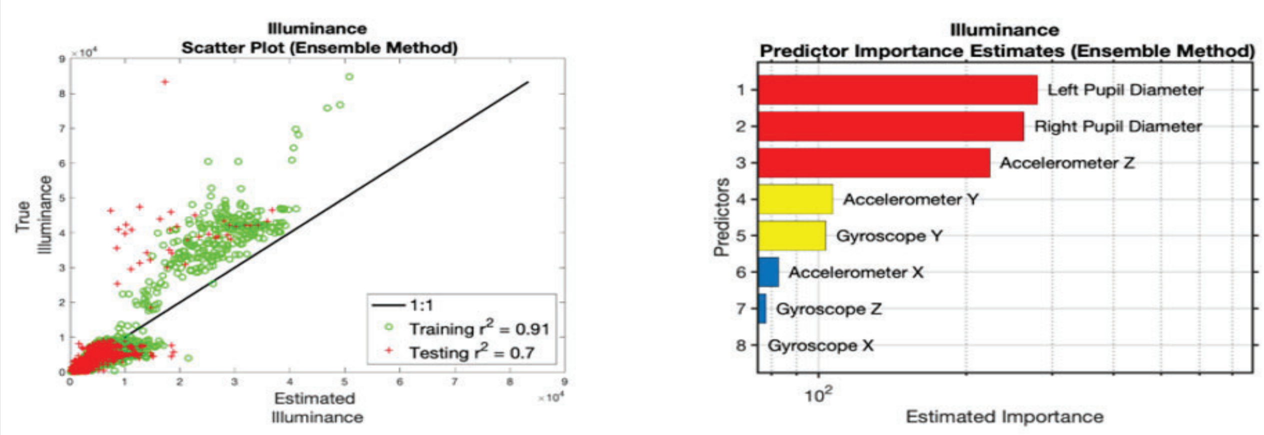


Figure 9: Plots for the illuminance model.

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).

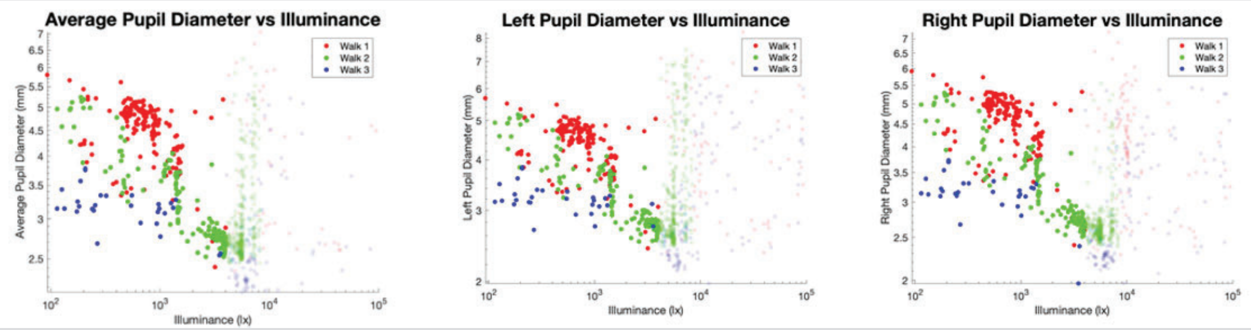


Figure 10: Log scale scatter plots of the pupil diameters vs illuminance. Data from walks 1, 2, and 3 are distinguished by the colors red, green, and blue, respectively. Data points with low opacity have illuminance values above 4000 lx. Note below the 4000 lx mark the variables tend to have an inverse relation- ship.

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)

(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 qual-ity, 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 per-formed. Cyclist performance can be assessed via bicycle speed and biometric data. Biometrics such as Electroencephalography (EEG), Heart Rate (ECG), Gal-Vanic 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].

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 effort 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 Scientific 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 fine 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

This work is licensed under Creative Commons Attribution 4.0 License

Submission Link: https://biomedres.us/submit-manuscript.php

Assets of Publishing with us

- Global archiving of articles
- Immediate, unrestricted online access
- Rigorous Peer Review Process
- Authors Retain Copyrights
- Unique DOI for all articles

https://biomedres.us/

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**[†]

* Correspondence: David.Lary@utdallas.edu

† Hanson Center for Space Sciences, University of Texas at Dallas, Richardson TX 75080, USA

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 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 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.; Lary, M.D. Autonomous Learning of New Environments With a Robotic Team Employing Hyper-Spectral Remote Sensing, Comprehensive In-Situ Sensing and Machine Learning. *Sensors* **2021**, *1*, 0. https://dx.doi.org/10.3390/s1010000

Received:
Accepted:
Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Submitted to *Sensors* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

This paper describes a robotic team that can rapidly learn new environments. The system described here demonstrates a flexible paradigm that is easily scalable to multi-robot, multi-sensor autonomous teams. A case study is described for the rapid characterisation of the aquatic environment.

The aquatic environment was chosen, as it includes extra challenges with regards to 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 [1–4]. So being able to rapidly survey such areas to guide clean-up operations is of considerable use. It is also of use in

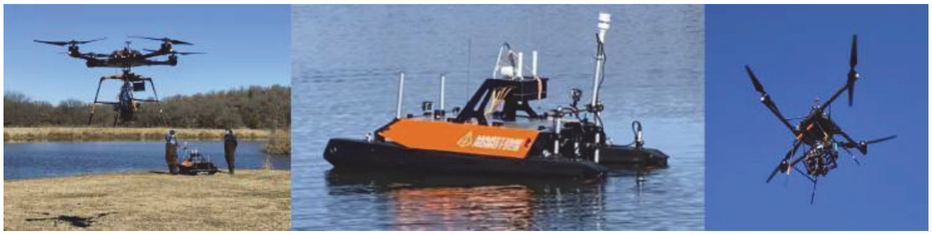


Figure 1. Photographs of the robot team during a Fall 2020 deployment in North Texas.

a wide variety of contexts, from general environmental surveys, to studying harmful algal blooms, to the clean-up operations after natural disasters, such as huricanes, etc.

In the example 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 on-board Machine Learning (ML) capability. Figure 1 shows photographs of the robot team during a December 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 paradigm demonstrated can reduce the time and cost of producing new remote sensing data products, while increasing functionality and data quality and providing new real-time automated calibration/validation capabilities.

The approach also provides enhanced capabilities for real-time on-board 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. A key element is the use of embedded machine learning, so we will refer to the approach as Rapid Embedded Prototyping for Advanced Applications (REPAA).

1.1. 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 391-1,011 nm. A hyper-cube is a three-dimensional dataset consisting of a stack of two-dimensional image layers each for a different wavelength. So for each pixel in the image we have a multi-wavelength spectra (spectral signature). This is shown schematically in the lower left of Figure 2. On the right we see a conventional RGB color images with only three bands, images for red, green and blue wavelengths.

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 down-welling 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.

An example reflectivity hyper-spectral data cube collected during a robot team deployment in North Texas during November 2020 is shown in Figure 3b. 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, that 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 1,011 nm.

2. Materials and Methods

All the data for the machine learning data product creation was collected in a coordinated automated manner using the autonomous robotic team. 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) autonomous boat was used. With a footprint of only 200 x 108 x 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 2 knots it has a duration of 20 hours 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) 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). 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). The airMAX Time Division Multiple Access (TDMA) protocol allows each client to send and receive data using pre-designated time slots managed by an intelligent AP controller. This

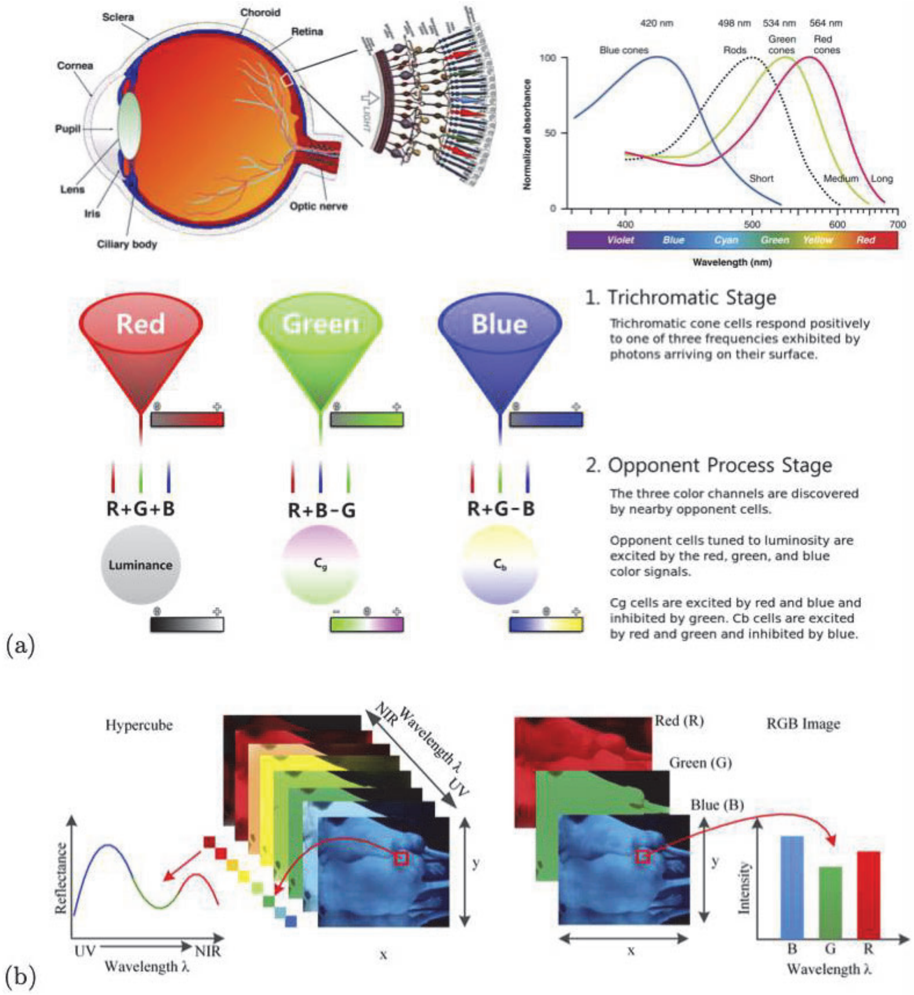

Sclera
Choroid
Retina
Cornea
Pupil
Lens
Iris
Ciliary body
Optic nerve
LIGHT
420 nm
498 nm
534 nm
564 nm
Blue cones
Rods
Green cones
Red cones
Normalized absorbance
100
50
0
Short
Medium
Long
400
500
600
700
Violet
Blue
Cyan
Green
Yellow
Red
Wavelength (nm)
Red
Green
Blue
1. Trichromatic Stage
Trichromatic cone cells respond positively to one of three frequencies exhibited by photons arriving on their surface.
2. Opponent Process Stage
The three color channels are discovered by nearby opponent cells.
Opponent cells tuned to luminosity are excited by the red, green, and blue color signals.
Cg cells are excited by red and blue and inhibited by green. Cb cells are excited by red and green and inhibited by blue.
R+G+B
R+B−G
R+G−B
Luminance
C_g
C_b
(a)
Hypercube
NIR
Wavelength λ
UV
Reflectance
Wavelength λ
x
y
Red (R)
Green (G)
Blue (B)
RGB Image
Intensity
B
G
R
(b)

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.

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).

This network also includes a local Synology network-attached storage (NAS) (https://www.synology.com) device in the robot team control trailer, which in real-time syncs the data 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/). Three Eureka Manta-40 multi-probes (https://www.waterprobes.com/multiprobes-and-sondes-for-monitori), a Sequoia Scientific LISST-ABS acoustic backscatter sediment sensor (https://www.sequoiasci.com/product/lisst-abs/), and a Airmar Technology Corporation 220WX ultra-sonic weather monitoring sensor (https://www.airmar.com/weather-description.html?id=153).

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) 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-

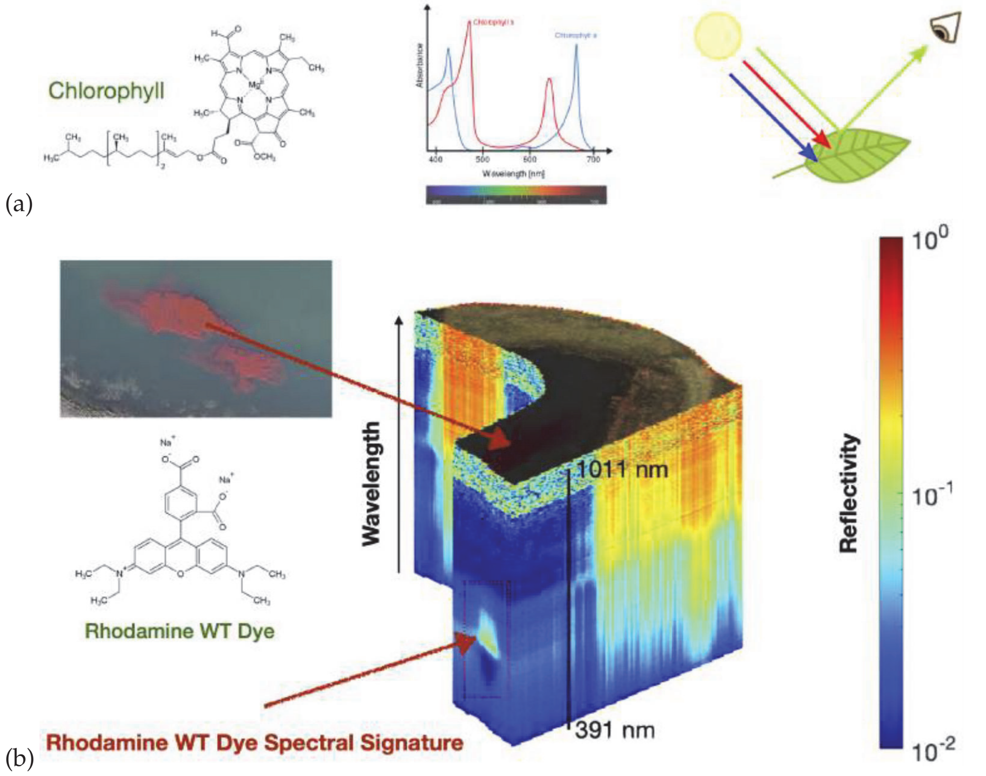


Figure 3. Panel (a) Chemicals absorb light in a characteristic way. Their absorption spectra is a function of their chemical structure. For every pixel we measure an entire spectrum with a hyper-spectral camera so we can identify chemicals within the scene. Panel (b) shows an example hyper-spectral data cube collected in North Texas on November 23, 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 463 stacked layers below show the reflectivity (on a log-scale) for each wavelength band between 391 and 1,011 nm.

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/) 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 seconds 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) made with aircraft grade aluminum and carbon fiber to carry a Resonon Visible+Near-Infrared (VNIR) Pika XC2 (https://resonon.com/Pika-XC2) hyper-spectral camera (391–1,011 nm) with a Schneider Xenoplan 1.4/17 mm lens, and a FLIR Duo Pro R, (640x512, 25 mm, 30 Hz) combining a high resolution, radiometric thermal imager, 4K color camera, and a full suite of on-board sensors (https://www.flir.com/products/duo-pro-r/). 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 are written in real time to the solid-state disk (SSD) attached to the Resonon Pika XC2. To facilitate the real-time processing of these files 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. These hyper-spectral data cubes provide a visible and near infrared spectrum (391–1,011 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 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 (391–1,011 nm) provided by the hyper-spectral data cubes collected by the aerial-vehicle. This data is 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 391–1,011 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 [5–22].

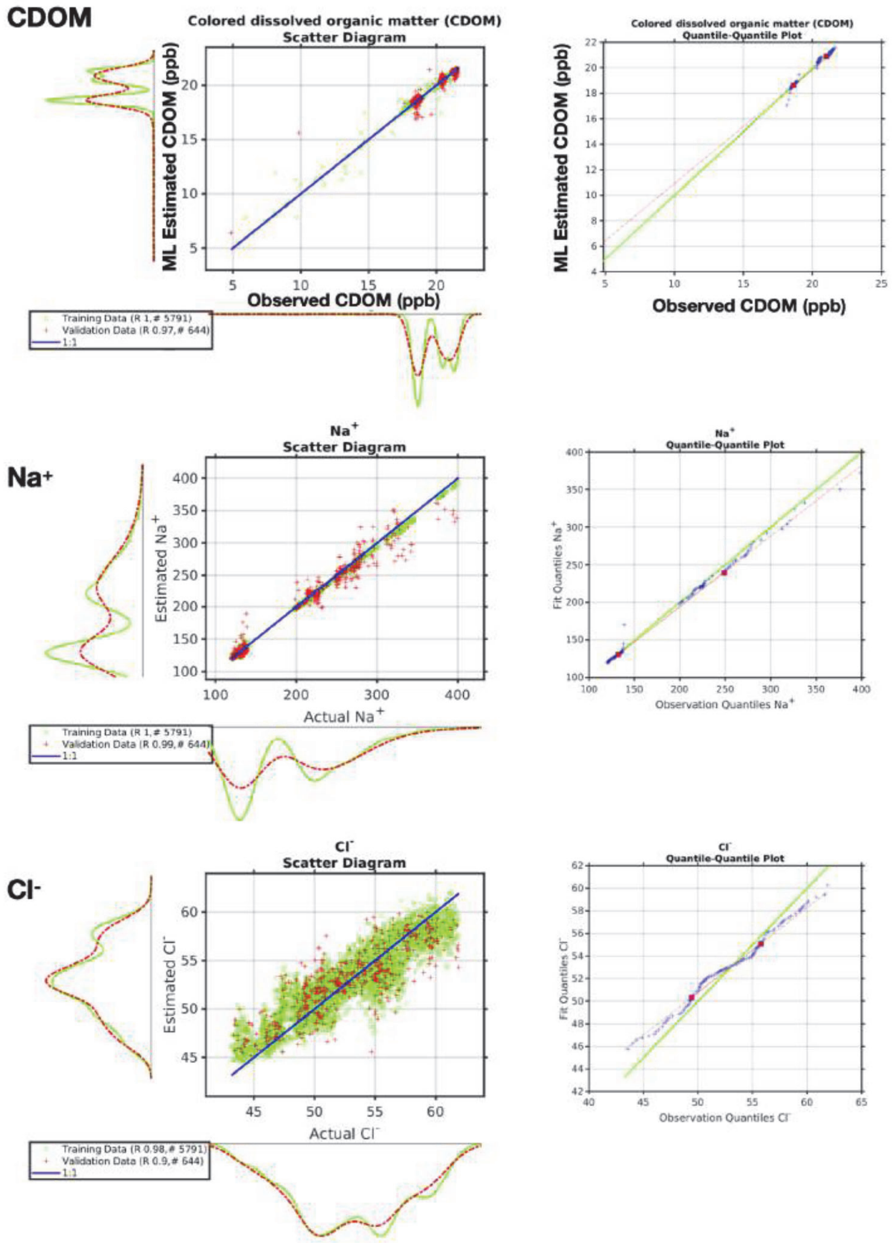
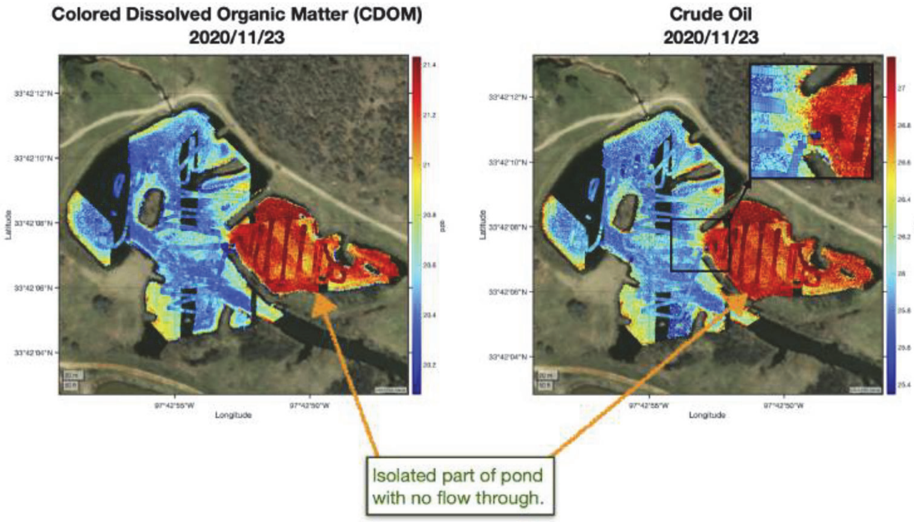


Figure 4. Machine learning performance quantified by both scatter diagrams and quantile-quantile plots utilizing data collected autonomously by the robot team during three exercises during November and December 2020 in North Texas. The three examples shown here are for CDOM, Na^{+} and Cl^{-}. The scatter 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.

3. Results

Over a period of just a few minutes we acquire thousands of training data points. This training data allows our machine learning algorithms to rapidly learn by example. The machine learning fit used here is a gaussian process regression [23] with


Colored Dissolved Organic Matter (CDOM) 2020/11/23

Crude Oil 2020/11/23

Isolated part of pond with no flow through.

Figure 5. 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.

Figure 4 shows an example of the colored dissolved organic mater (CDOM) data collected autonomously by the robot team on November 23, 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 5 shows maps of the CDOM and crude oil concentration estimated 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 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. We note that there is good agreement between the machine learning estimate and the actual in-situ boat observations.

4. Discussion

4.1. 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). The cost of making the measurements of ocean composition can be substantial because it involves significant ship time as well as a large support team. Secondly, since 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 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 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 on-board 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. [25] 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 oceanographic campaigns 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 6, is compared with the approach used in this study, 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).

4.2. 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 costs less. In summary, our robotic platform can address the issue of small scale variability encountered across a satellite pixel. These capabilities assist continuing validation/quality control and can help optimize the waveband selection for future satellite instruments and missions.

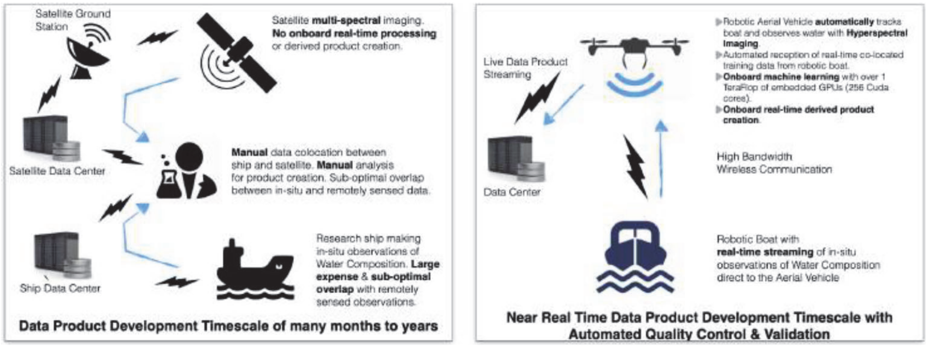

Satellite Ground Station
Satellite **multi-spectral** imaging. **No onboard real-time processing** or derived product creation.
Satellite Data Center
Manual data colocation between ship and satellite. **Manual** analysis for product creation. Sub-optimal overlap between in-situ and remotely sensed data.
Ship Data Center
Research ship making in-situ observations of Water Composition. **Large expense** & **sub-optimal overlap** with remotely sensed observations.
Data Product Development Timescale of many months to years
Live Data Product Streaming
Data Center
Robotic Aerial Vehicle **automatically** tracks boat and observes water with **Hyperspectral Imaging**.
Automated reception of real-time co-located training data from robotic boat.
Onboard machine learning with over 1 TeraFlop of embedded GPUs (256 Cuda cores).
Onboard real-time derived product creation.
High Bandwidth Wireless Communication
Robotic Boat with **real-time streaming** of in-situ observations of Water Composition direct to the Aerial Vehicle
Near Real Time Data Product Development Timescale with Automated Quality Control & Validation

Figure 6. Schematics illustrating the traditional approach to creating remote sensing data products (left) and that used in this study (right).

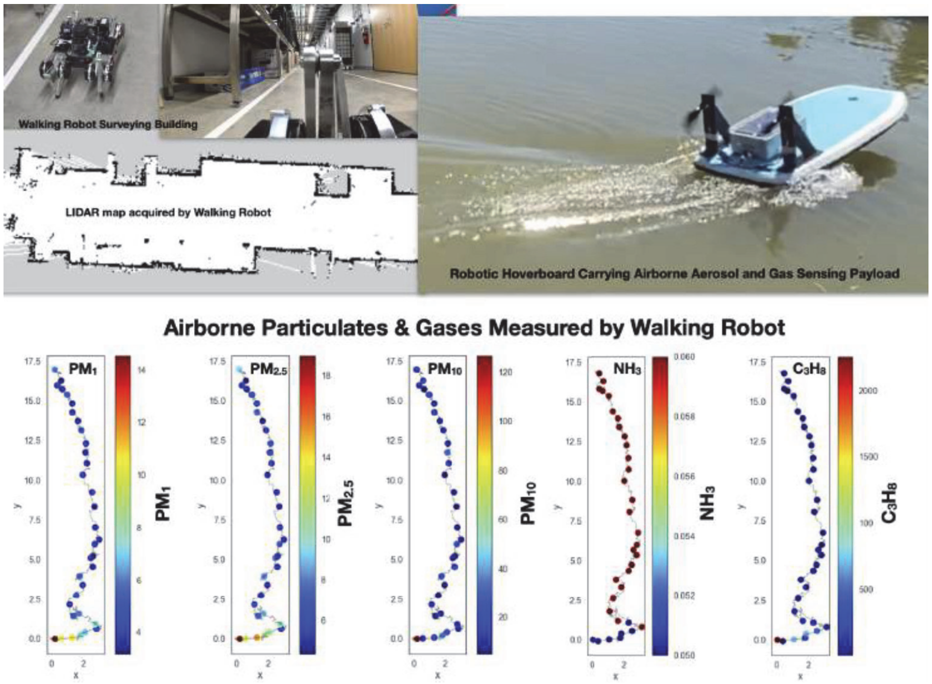


Figure 7. Photographs of the smaller walking robot (from Ghost Robotics) and a robotic hoverboard (conceived and built by Aaron Barbosa) that for illustrative purposes both carried exactly the same payload of sensors measuring the size spectrum of airborne particulates in the size range 0.3–43 microns 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.

4.3. Reducing Latency for Product Delivery as well as Mission Risk, Cost, Weight and Size

Utilizing new embedded *on-board processing* (1 TeraFlop weighing just 88 g with a size of only 87 mm x 50 mm) for real-time on-board 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, using entirely commercial off the shelf components and low cost aerial vehicles, with all software made Open Source.

4.4. 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 required for the downlink of the hyperspectral datacubes. So 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 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 come for this to be the standard paradigm for earth observation as well.

4.5. Smaller Robots

There is also value in smaller robots that are easy to transport by a single individual. Figure 7 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. Both the walking robot and the robotic hover-board 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.

5. 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 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. 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, each measuring the full size spectrum of airborne particulates in the size range 0.3–43 microns and the abundance of a selection of gases, significant small scale spatial variability with evident in these hyper-localized observations.

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, 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.

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.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Acknowledgments: Mr. Don MacLaughlin, Mr. 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. Prof. 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.

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. Fingas, M.F.; Brown, C.E. Review of oil spill remote sensing. *Spill Science & Technology Bulletin* **1997**, *4*, 199 – 208. The Second International Symposium on Oil Spills, doi:http://dx.doi.org/10.1016/S1353-2561(98)00023-1.
2. Fingas, M. *Oil Spill Science and Technology*; Gulf Professional Publishing, 2010.
3. Liu, Y.; MacFadyen, A.; Ji, Z.; Weisberg, R. *Monitoring and Modeling the Deepwater Horizon Oil Spill: A Record Breaking Enterprise*; Geophysical Monograph Series, Wiley, 2013.
4. Cornwall, W. Deepwater Horizon: After the oil. *Science* **2015**, *348*, 22–29, [http://www.sciencemag.org/content/348/6230/22.full.pdf]. doi:10.1126/science.348.6230.22.
5. 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. *International Journal of Remote Sensing* **2008**, *29*, 7141–7158.
6. Lary, D.; Aulov, O. Space-based measurements of HCl: Intercomparison and historical context. *Journal of Geophysical Research: Atmospheres* **2008**, *113*.
7. Lary, D.J.; Remer, L.; MacNeill, D.; Roscoe, B.; Paradise, S. Machine learning and bias correction of MODIS aerosol optical depth. *IEEE Geoscience and Remote Sensing Letters* **2009**, *6*, 694–698.
8. Lary, D.; Waugh, D.; Douglass, A.; Stolarski, R.; Newman, P.; Mussa, H. Variations in stratospheric inorganic chlorine between 1991 and 2006. *Geophysical Research Letters* **2007**, *34*.
9. Lary, D.; Müller, M.; Mussa, H. Using neural networks to describe tracer correlations **2004**.
10. Lary, D.J. *Artificial intelligence in geoscience and remote sensing*; INTECH Open Access Publisher, 2010.
11. Malakar, N.K.; Knuth, K.H.; Lary, D.J., Maximum Joint Entropy and Information-Based Collaboration of Automated Learning Machines. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*; Goyal, P.; Giffin, A.; Knuth, K.H.; Vrscay, E., Eds.; 2012; Vol. 1443, *AIP Conference Proceedings*, pp. 230–237. doi:10.1063/1.3703640.
12. 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 (PM2.5). *Geospatial health* **2014**, *8*, 611–630.
13. Lary, D.; Lary, T.; Sattler, B. Using Machine Learning to Estimate Global PM2. 5 for Environmental Health Studies. *Environmental health insights* **2015**, *9*, 41.
14. Lary, D.J.; Alavi, A.H.; Gandomi, A.H.; Walker, A.L. Machine learning in geosciences and remote sensing. *Geoscience Frontiers* **2016**, *7*, 3–10.
15. 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. *Atmospheric Environment* **2016**, *128*, 53–64.
16. 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. *Environmental Health Insights* **2017**, *11*, 1–10.
17. Nathan, B.J.; Lary, D.J. Combining Domain Filling with a Self-Organizing Map to Analyze Multi-Species Hydrocarbon Signatures on a Regional Scale. *Environmental Modeling and Assessment* **2019**, *191*.
18. Lary, D.J.; Zewdie, G.K.; Liu, X.; Wu, D.; Levetin, E.; Allee, R.J.; Malakar, N.; Walker, A.; Mussa, H.; Mannino, A.; others. Machine Learning Applications for Earth Observation. In *Earth Observation Open Science and Innovation. ISSI Scientific Report Series*; Springer, 2018; Vol. 15, pp. 165–218.
19. Lary, M.A.; Allsop, L.; Lary, D.J. Using Machine Learning to Examine the Relationship Between Asthma and Absenteeism. *Environmental Modeling and Assessment* **2019**, *191*.
20. 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. *International journal of environmental research and public health* **2019**, *16*, 1992.
21. 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. *Environmental Monitoring and Assessment* **2019**, *191*, 418.

22. 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.
23. Rasmussen, C.E. Gaussian processes in machine learning. Summer School on Machine Learning. Springer, 2003, pp. 63–71.
24. Nocedal, J.; Wright, S. *Numerical optimization*; Springer Science & Business Media, 2006.
25. Aurin, D.; Maninno, A.; Lary, D.J. Remote Sensing of CDOM, CDOM Spectral Slope, and Dissolved Organic Carbon in the Global Ocean. *Applied Sciences* **2018**, *8*, 2434.

remote sensing

MDPI

Article

Cloud Detection Using an Ensemble of Pixel-Based Machine Learning Models Incorporating Unsupervised Classification

Xiaohe Yu [1,*] and David J. Lary [2]

[1] Geospatial Information Sciences, The University of Texas at Dallas, Richardson, TX 75080, USA
[2] Hanson Center for Space Science, The University of Texas at Dallas, Richardson, TX 75080, USA; David.Lary@utdallas.edu
* Correspondence: xhyu66@gmail.com or xxy160430@utdallas.edu

Abstract: Remote sensing imagery, such as that provided by the United States Geological Survey (USGS) Landsat satellites, has been widely used to study environmental protection, hazard analysis, and urban planning for decades. Clouds are a constant challenge for such imagery and, if not handled correctly, can cause a variety of issues for a wide range of remote sensing analyses. Typically, cloud mask algorithms use the entire image; in this study we present an ensemble of different pixel-based approaches to cloud pixel modeling. Based on four training subsets with a selection of different input features, 12 machine learning models were created. We evaluated these models using the cropped LC8-Biome cloud validation dataset. As a comparison, Fmask was also applied to the cropped scene Biome dataset. One goal of this research is to explore a machine learning modeling approach that uses as small a training data sample as possible but still provides an accurate model. Overall, the model trained on the sample subset (1.3% of the total training samples) that includes unsupervised Self-Organizing Map classification results as an input feature has the best performance. The approach achieves 98.57% overall accuracy, 1.18% cloud omission error, and 0.93% cloud commission error on the 88 cropped test images. By comparison to Fmask 4.0, this model improves the accuracy by 10.12% and reduces the cloud omission error by 6.39%. Furthermore, using an additional eight independent validation images that were not sampled in model training, the model trained on the second largest subset with an additional five features has the highest overall accuracy at 86.35%, with 12.48% cloud omission error and 7.96% cloud commission error. This model's overall correctness increased by 3.26%, and the cloud omission error decreased by 1.28% compared to Fmask 4.0. The machine learning cloud classification models discussed in this paper could achieve very good performance utilizing only a small portion of the total training pixels available. We showed that a pixel-based cloud classification model, and that as each scene obviously has unique spectral characteristics, and having a small portion of example pixels from each of the sub-regions in a scene can improve the model accuracy significantly.

Keywords: landsat 8; machine learning; cloud detection; ensemble approaches; self organizing maps (SOM); NDSI; NDVI; whitness; HOT

check for updates

Citation: Yu, X.; Lary, D.J. Cloud Detection Using an Ensemble of Pixel-Based Machine Learning Models Incorporating Unsupervised Classification. *Remote Sens.* **2021**, *13*, 3289. https://doi.org/10.3390/rs13163289

Academic Editor: Filomena Romano

Received: 16 July 2021
Accepted: 17 August 2021
Published: 20 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Remote sensing imagery, such as that provided by the United States Geological Survey (USGS) Landsat satellites, has been widely used to study environmental protection, hazard analysis, and urban planning for decades. The usefulness and applications of remote sensing imagery continue to expand as more image-based models and algorithms have emerged so that we can derive more knowledge and information from satellite images. Due to the ubiquity of clouds, cloud pixels are a persistent presence in such imagery, especially in tropical areas. A study estimates that about 67% of the earth's surface is typically covered by cloud based on satellite data from July 2002 to April 2015 [1]. The presence of cloud has a serious impact on the use of remote sensing images. Cloud areas appear as extremely bright pixels in the images. These pixels can cause issues in various remote sensing imagery analyses, including incorrect land surface classification, inaccurate atmosphere correction, low quality Aerosol Optical Depth (AOD) retrieval and false land surface change detection [2]. As a result, clouds are considered noise in most situations and are typically removed prior to further analysis, which makes cloud detection a crucial step for remote sensing image preprocessing.

Over the last few decades, a variety of methods have been developed for cloud detection. Let us briefly consider a few examples.

1.1. An Overview of Cloud Detection Approaches

Automated Cloud Cover Assessment (ACCA) [3], a scene specific cloud detection method was developed for Landsat 7. It employs two pass through ETM+s to establish the the reflective and thermal features of the cloud and non-cloud area in a scene, and then to identify the cloud in the rest area of the whole scene. This approach experiences difficulty in identifying cloud areas with snow and brightly illuminated desert areas.

Zhu et al. [4] proposed Function of Mask (Fmask) which is an objected-based cloud and cloud shadow approach for Landsat images, which has a 96.41% reported accuracy. The author of Fmask further improved Fmask in terms of increasing the performance for Landsat 4–7, making it suitable for Landsat 8 and Sentinel 2 imagery [5].

Foga et al. [6] compared the performance of ACCA, LEDAPS CCA, and CFmask (C version of Fmask) on three cloud validation dataset including IRISH, SPARCS, and Biome. Among these three algorithms, CFmask is reported with the best overall accuracy.

Hughes et al. [7] proposed a machine learning approach for automated cloud and cloud shadow detection by using neural network and spatial post-processing techniques, which achieves lower cloud shadow omission error and cloud commission error compared to Fmask.

A multi-feature combined (MFC) approach is proposed for the Chinese GaoFren1 cloud detection by using spectral features in combination with geometric and texture features, which has a 96.8% accuracy [8].

All these aforementioned approaches are single temporal approaches, which only require one scene for implementation. In contrast to the single temporal approach, a multi-temporal approach (Tmask) is proposed for cloud, shadow, and snow by using multiple images at the same location [9]. It generates a time series model which is used to predict the TOA reflectance surface. These surfaces are then compared with Landsat images to differentiate clouds, shadows, and snow. However, this approach requires at least 15 clear observations in each pixel to generate a robust time series model, which makes it less applicable in places that have long been covered by snow or cloud.

Candra et al. [10] proposed an automated cloud and cloud shadow detection method by using multi temporal Lansat8 images, which is named MCM. This approach makes use of the reflectance differences between two images at the same location to identify cloud and cloud shadows, which is especially effective in tropical areas.

The use of deep learning techniques including CNN, RNN and GCN have recently garnered much attention for remote sensing image classification tasks because they are capable of extracting high-level features from images. Zhu et al. [11] reviewed the major advances of deep learning in remote sensing. Xie et al. [12] proposed a multilevel cloud detection method based on simple linear iterative clustering (SLIC) and a deep convolutions neural network (CNN). This method achieves a better result compared with a scene learning-based approach proposed by Zhenyu [13] and progressive refinement scheme approach proposed by Zhang et al. [14]. Authors of [15] proposed a cloud detection method (MSCN) based on Fully Convectional Networks (FCN) [16] by fusing multi-scale convolutional features for cloud detection, which is effective in snow and areas covered by non-cloud bright objects.

Another CNN-based cloud detection method is trained on high resolution WV-2 satellite images, which not rely on SWIR or IR bands and can be applied to Sentinel imagery [17]. Zi et al. [18] proposed a novel cloud detection method for Landsat 8 images by using Simple Linear Iterative Cluster (SLIC), PCA Network (PCANet), Support Vector Machine (SVM), and Conditional Random Field (CRF). This approach combines statistical models, classical machine learning methods, and a deep learning network to generate a robust model for cloud detection and achieves an accurate result. Yang et al. [19] proposed a CNN-Based cloud detection method by using thumbnails of remote sensing images instead of the original remote sensing images. To handle the coarse resolution of thumbnail images, a cloud detection neural network feature pyramid module, and boundary refinement block techniques are employed to generate accurate cloud prediction results. This work has been further extend by Guo [20] for cloud and snow coexistence scenarios by proposing a new model DSnetV2.

In addition to cloud detection, deep learning-based techniques have been extensively applied to remote sensing image classification problems, especially for hyper-spectral images. Graph convolutional networks (GCNs) are a new emerging network architecture that can handle and model long-range spatial relations. Shahraki and Prasad [21] proposed a cascade framework of 1-D CNNs and GCNs for a hyper-spectral classification problem. Qin et al. [22] extended the GCNs by considering spatial and spectral neighbors. Pu et al. [23] proposed a localized graph convolutional filtering-based GCNs method for hyper-spectral image classification. Traditional GCNs are computationally expensive because the spatial matrices are constructed. Hong et al. [24] showed that miniGCNS can be trained in minibatch fashion for classification problems. The miniGCNs are more robust, and are capable of handling out-of-samples with lower computation cost compared to traditional GCNs.

Based on the data required, these cloud detection methods can be divided into single temporal and multi-temporal approaches. The single temporal approach seeks to identify cloud pixels based on imagery at a single time, while a multi-temporal approach makes use of imagery from multiple comparable timeframes for a same area to identify cloud pixels by comparing the pixel differences between cloud free images and cloudy ones [25]. Depending on the algorithm used, cloud detection can be categorized into a classical algorithm-based approach and machine learning approach [26]. The classical algorithm refers to methods which have specific steps to be followed for input imagery and to generate the output mask like the FMask [4]. On the other hand, machine learning approaches take the advantage of existing data and learn from a training set without human interference to generate an output [18].

Cloud detection is still challenging in these aspects. First, cloud pixels are hard to identify from "bright" areas such as snow by traditional rule-based approaches. The multi-temporal approach requires more than one image at the same location, which is less applicable for low temporal resolution satellites, such as Landsat. While deep learning-based approaches generally enable better cloud detection accuracy, they require a high performance GPU which may not be available. Some other approaches require additional information in combination with the spectral features to improve the performance, which requires extra labor efforts.

1.2. A Pixel-Based Approach Using an Ensemble of Learners

This paper has four distinct goals. (1) Propose a cloud detection modeling approach for cloud detection by only using the 10 wavelength bands available at a *single pixel* as an information source without the need of any other ancillary data. (2) Engineer important predictor features that could increase the model accuracy. (3) Investigate the influence of the training sample size on the model accuracy, thus finding the smallest possible training sample size that could balance the training time and machine learning model accuracy. (4) Explore the importance rank of predictors and the optimal hyper-parameter settings for cloud mask prediction.

In contrast to the whole image-based approaches just described, and to mitigate the challenges just enumerated, in this study we have taken a pixel-based approach that only requires a single image. We have used an ensemble machine learning approach that has been tested with Landsat 8 imagery.

Compared with the multi-temporal approaches [25], this single-temporal approach is more feasible because of the mono-temporal feature of Landsat data in nature. Unlike other classical approaches that need auxiliary data [4], this approach only requires the information on the 10 wavelength bands for a single pixel of Landsat 8 imagery. In comparison to machine learning approaches under a deep learning frame work, the computational facility is not as demanding as deep learning, so a GPU is not needed in this research. This proposed machine learning model uses an ensemble approach which simultaneously employs multiple decision tree learners for cloud detection. The input parameters are tuned in two ways. On the one hand we optionally include unsupervised classification results from the application of a self-organizing map (SOM) as one of the input features for the ensemble model training. On the other hand, 5 indices that are calculated from the 10 wavelength bands signal are included as input features for model training.

Four distinct training subsets were generated from the Biome cloud validation dataset for model training [6]. Models are generated based on different sets of input parameters and different training samples. Then, their performance is compared against Fmask 4.0.

2. Materials and Methods

2.1. Data Sources

This study uses the L8 cloud cover assessment validation dataset as the source for model training and validation [6,27]. The L8 Biome dataset includes 96 Landsat 8 images sampled across the world, and with a manually generated cloud mask for each image. The name Biome is given because the target scenes are sampled based on biome types. The path/row of the scenes covering the land are sampled across the world on the basis of biome types which include urban, forest, shrubland, grass/cropland, snow/ice, wetlands, and water. In order to create a heterogeneous dataset, path/rows with obviously clustering patterns are discarded and will be re-sampled from their biome types. Then scenes within the final set of path/rows will be labeled as clear, mid-cloudy, and cloudy manually based on the cloud coverage percentage within each scene. A total of 96 scenes are prepared covering 8 biome types in 3 cloud coverage levels as shown in Figure 1.

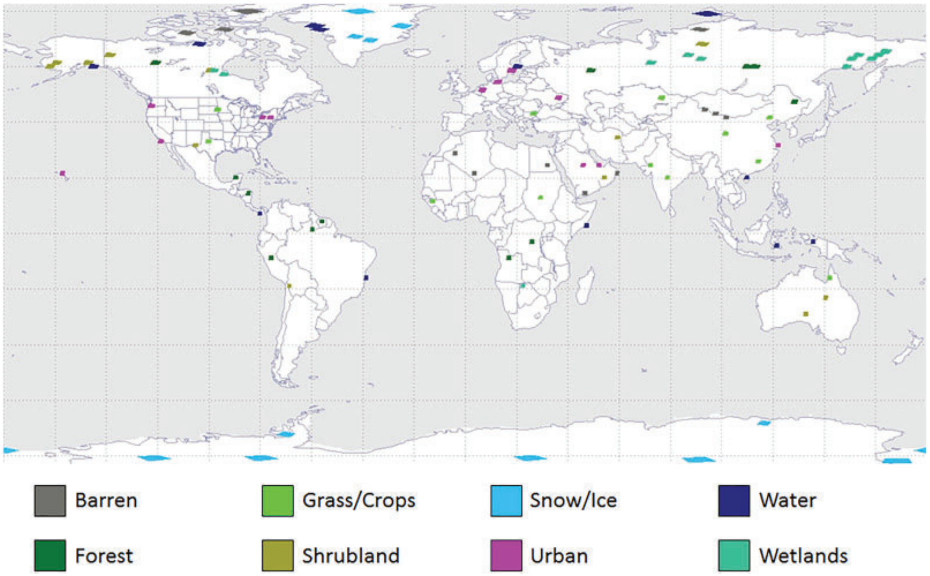


Figure 1. Global distribution of the 96 unique Landsat 8 Cloud Cover Assessment scenes, adopted from [6].

2.2. Data Pre-Processing: The Intra-Group and the Ultra-Group

Validation masks and all spectrum band layers from the Biome dataset are cropped to 4000 × 4000 pixels for model training and validation (Figure 2). These 96 cropped Biome images are then divided into two groups, the intra-group and the ultra-group. The intra-group includes 88 images from which we sub-sample to provide our training pixels. The name "intra-group" rather than "training-group" was given, because only a tiny portion (up to 1.3%) of the intra-group are used as the model training source to show the effectiveness of our pixel based approach. Therefore, more than 98% of the data in the intra-group are "new" to the models, which also makes the intra-group qualified for model performance evaluation. On the other hand, the ultra-group contain 8 images that our models have never seen, which means the ultra-group are 100% "new" to the models. Images in both groups will be used for model performance evaluation.

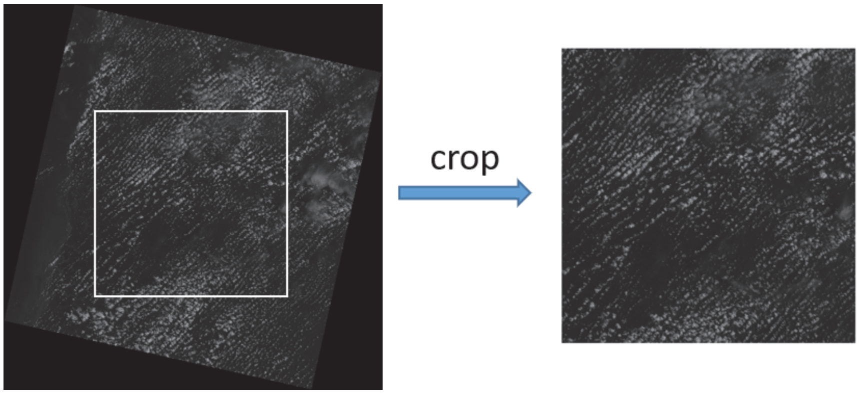


Figure 2. Lansat8 scenes are centre-cropped into 4000 × 4000 pixel images.

For each biome type, eleven cropped images are selected as the intra-group, and we hold the one remaining image as the ultra-group (See Table 1). Then the 10 wavelength bands of each scene are stacked as a TIFF file for model training and validation. However, band 8 has a 15 m resolution which is higher than other bands. We downgrade the resolution for band 8 to make the imagery resolution the same as the other bands.

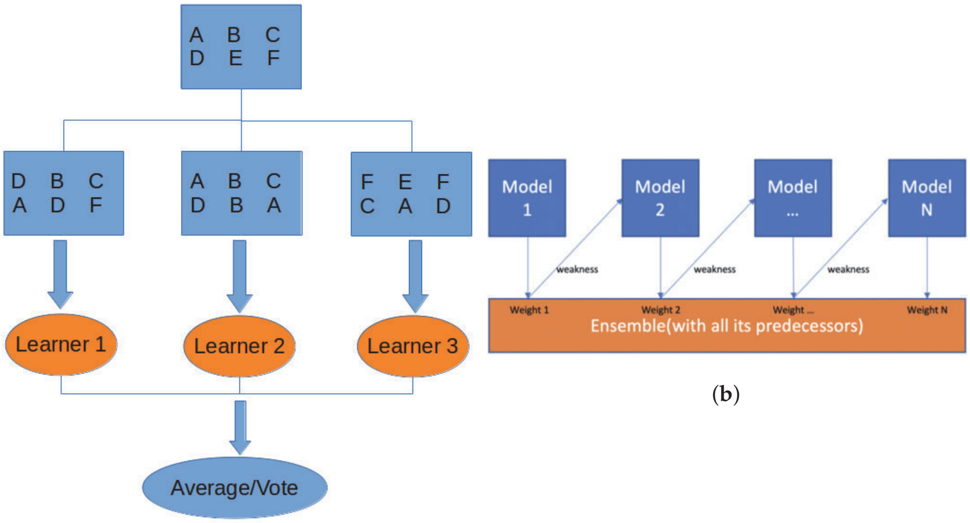
Table 1. Number of Scenes included in the Intra-Group and the Ultra-Group.

Intra-Group	Ultra-Group	Land Types
11	1	Barren
11	1	Forest
11	1	Grass/Crops
11	1	Shrubland
11	1	Snow/Ice
11	1	Water
11	1	Urban
11	1	Wetlands
88	8	Total

2.3. Ensemble Machine Learning Classification Approach

Ensemble learning is a machine learning paradigm where multiple weak learners are combined to improve the training performance. Various types of learner aggregation can be employed, including Bootstrap Aggregation (Bagging), Boosting and random space. In this paper, all three of these tree ensemble approaches are used, and the particular approach that will be employed in the tree-based ensemble model will be determined at the Bayesian Optimization stage.

Bagging aggregates trained weak learners by creating many bootstrap replicas and training each weak learner on one replica. Typically, the number of bootstrap replicas can vary from just a few up to several hundred. Each replica is generated by drawing N samples out of N observations with replacement (see Figure 3a). Drawing samples with replacement omits an average of 37% of the total observations which decreases the correlation between each weak learner to avoid over-fitting. Each weak learner is trained on a single replica by randomly extracting features for each split node which is called the random forest technique. Then the ensemble model makes a prediction on new data by taking the most voted predictions from individual learners for use in the classification.



(**a**)

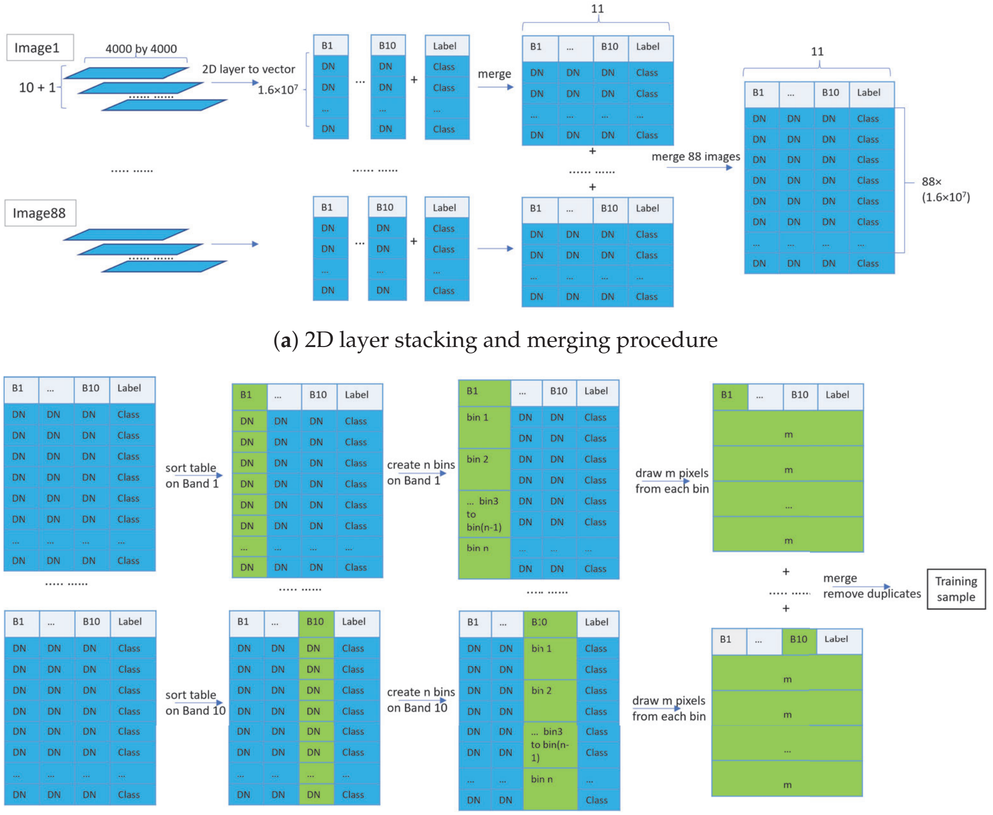
(**b**)

Figure 3. Schematics illustrating the various ensemble model approaches. (**a**) Bagging, (**b**) Boosting.

Boosting trains learners sequentially (see Figure 3b). It maintains a weight distribution for all training observations and each observation is assigned a weight indicating its importance. By decreasing the weight for correctly classified observations and increasing the weight for misclassified observations at each round, trained learners are able to focus more on hard observations that have been misclassified by previous learners. Distinct from bagging which generates replicas for individual learner training, boosting trains all models with the same dataset but with different weight. Instead of taking the most votes of the predictions from individual learners, boosting combines the predictions from individual learners with weights. As a result, boosting is able to make individual learners focus more on hard observation points round by round, thus to obtain a premium result from weak learners. In this paper, AdaBoost, RUSboost, and LSBoost are candidate boosting methods employed for our ensemble models.

2.4. Representative Sampling

There are a total of 88 4000 × 4000 pixel images in the intra-group, a total of 1.408 billion pixels. A small portion of the samples will be drawn from the intra-group and used for model training. Generally, the more training data we use, the better machine learning model performance we achieve. However, for the task of cloud detection, the cloud label is very labor intensive to generate and is not widely available. One goal of this research is to explore a machine learning modeling approach that uses as small a fraction of the intra-group dataset as possible but can still fit the data well. In order to compare the performance of models trained on different sizes of training subsets, a representative random sampling approach is proposed to sample four subsets from the 88 cropped images. Each of the 88 images has 10 wavelength bands. After stacking the 10 wavelength bands together, each stacked image becomes a 4000 × 4000 × 10 matrix. Then a validation label is attached to the stacked matrix thus forms a 4000 × 4000 × 11 matrix for each image. These 3-dimensional images are converted to 2-dimension image tables, which are then concatenated along the horizontal direction and form an image pool table (Figure 4a). The next step is to draw representative samples from this image pool table (Figure 4b). One column of the image pool table is sorted and divided into n bins. These bins have an uniform width covering the range of pixel values of the column, which could potentially reveal the distribution of values in the column. Then m samples are randomly drawn from each bin of the column and this sampling process is repeated for all columns in the image pool table. Duplicated rows could be drawn when repeating the sampling process across each column, which could lead to an unbalanced sampling subset. Any duplicated samples are removed after the sampling process complete, and a sample subset that covers the value range of each column is obtained without any duplication. The bin number n and the sample number m are adjusted to obtain different training samples from the whole dataset. In this research, 4 training samples subsets are generated for model training (see Table 2).



(**a**) 2D layer stacking and merging procedure

(**b**) Representative sampling procedure

Figure 4. Flow chart illustrating the image stacking and representative sampling procedures. In the stacking and merging procedure (**a**), the 10 wavelength bands as well as the label mask for each of the 88 images are stacked as an image table. Then, the 88 image tables are merged into one image pool table. The B1–B10 and Label in the headers represent the spectral band number and the pixels class labels. DN in tables represent the digital value. In the representative sampling procedure (**b**), the image pool table is sorted on a certain column (in green), then n bins are selected from the sorted column and m pixels are drawn from each bin. A final training sample is generated after repeating this process over all columns.

Table 2. Representative Training Samples. n represents the number of bins, m represents the number of pixels sampled in each bin, and the Total represents the total pixels sampled for each subset.

Subset Name	Bins (n)	Observations in Bin (m)	Total	Percentage
Sample 1	256	100	185,841	0.01%
Sample 2	512	200	714,972	0.05%
Sample 3	1024	300	2,063,945	0.15%
Sample 4	5120	600	18,722,275	1.30%

2.5. Feature Selection

2.5.1. Indices for Cloud Differentiating

A set of 5 indices is derived from the 10 wavelength bands and used as input features. These include $NDSI$, $NDVI$, $Whiteness$, HOT, and $B5/B6$ *ratio* (See Equations (1)–(4)). These indices are sensitive to cloud pixels and often have a threshold that are able to differentiate cloud pixels from others [4]. Instead of testing these indices with a threshold value, they are integrated as part of input parameters into the proposed machine learning model.

$$\begin{aligned} NDSI &= (Band\ 3 - Band\ 6)/(Band\ 3 + Band\ 6) \\ NDVI &= (Band\ 5 - Band\ 4)/(Band\ 5 + Band\ 4) \end{aligned} \tag{1}$$

$NDSI$ and $NDVI$ represent the normalized difference snow index and the normalized difference vegetation index, respectively. Both indices usually have a value near zero for clouds. Sometimes they could be larger for thin cloud over highly vegetated areas but should be less than 0.8.

$$\begin{aligned} MeanVis &= (Band\ 2 + Band\ 3 + Band\ 4)/3 \\ Whitness &= \sum_{i=2}^{4} \left| \frac{(Band\ i - MeanVis)}{MeanVis} \right| \end{aligned} \tag{2}$$

The Whiteness index was first proposed by Gomez-Chova et al. [28] for Environmental SATellite (ENVISAT) and was modified as in Equation (2) later [4] for Landsat satellite images. The modified whiteness is effective in excluding non-cloud pixels due to their “non-flat” reflectance feature compared to the “flat” reflectance of cloud pixels.

$$HOT = Band\ 2 - 0.5 \times Band\ 4 - 0.08 \tag{3}$$

The Haze Optimized Transformation (HOT) index was first proposed by Zhang et al. [29] and was modified by Zhu et al. [4], which is useful to separate haze and thin cloud from clear-sky pixels.

$$B5/B6\ Ratio = \frac{B5}{B6} \tag{4}$$

The $B5/B6$ ratio is able to separate “bright” rocks from cloud pixels as “bright” rocks usually have a higher *Band* 5 value than *Band* 4 while cloud has a higher *Band* 4 value than *Band* 5.

2.5.2. Self Organizing Map

A self-organizing map (SOM) is an unsupervised clustering method that uses a neural network to represent training data in a low-dimensional space [30]. The 10-dimensional input vectors are grouped into 100 clusters with 100 neurons. Each neuron has a weight vector which will be updated based on competitive learning approach at iterations. At each iteration, all neurons will be compared with a random selected training vector, and the one that is the closest in distance to the training point will be the winner and rewarded for moving closer to the training point by updating the weights. A neighborhood function is used to update the weights of the neighbors of a winning neuron in order to preserve the topological properties of the input space. Each cluster links to those image pixels sharing similar features. Although the SOM clustering results may not be able to directly differentiate cloud pixels from others pixels, it can provide useful information for the pixel type that can be utilized as an input feature for the training of the ensemble models. The SOM learning process can be represented by the neuron weight update process that is defined by (5).

$$\begin{aligned}\Delta W_{j,i} &= \eta(t) \cdot T_{j,I(x)}(t) \cdot (x_i - wi, j) \\ \eta(t) &= \eta_0 \exp\left(-\frac{t}{\tau_\eta}\right) \\ T_{j,I(x)}(t) &= \exp\left(-\frac{S^2_{j,I(x)}}{2\sigma(t)^2}\right) \\ S_{j,i} &= ||w_j - w_i|| \\ \sigma(t) &= \sigma_0 \exp\left(-\frac{t}{\tau_0}\right)\end{aligned} \tag{5}$$

where $\Delta W_{j,i}$ denotes the updated weight value, t is the epoch, i and j are neuron indices, and $I(x)$ is the winning neuron. $\eta(t)$ refers to the learning rate which controls the neuron learning speed. $T_{j,I(x)}(t)$ denotes the neighborhood function which defines the extent of neighbor neurons of a winning neuron. $S_{j,i}$ is the distance between neurons, and the $S_{j,i}$ is the neighborhood size.

2.6. Model Training and Hyper-Parameter Optimization

2.6.1. Model Training

A total of twelve tree-based ensemble models were trained on the four subsets sampled from the 88 cropped Landsat 8 images. Depending on the selection of input features, these machine learning models could be divided into three groups. Each model group consists of four models with the same specifications but trained on four different sizes of training samples.

The features used to train the first model group were the 10 wavelength bands from the 88 cropped images. For the second model group, in addition to the 10 wavelength bands, the 5 derived indices discussed in Equations (1)–(4) were also included as input features. The training process of the third model group consists of two steps. The first step is to build an SOM model on the training sample and use the SOM model to classify all the pixels from the training sample into clusters. The second step is to then include the SOM cluster label as an extra input feature in addition to the 10 wavelength bands. Then an ensemble learning model will be trained by employing the 10 wavelength bands and the SOM labels as the input features. The target variables for each subset are the pixel class which consists of cloud, clear, thin cloud, and cloud shadow. Models are summarized in Table 3 and the sample sizes are summarized in Table 2.

Table 3. Model names and features.

Model Name	Input Features	Training Subset
88Mdl_Sample 1	10 bands	Sample 1
88Mdl_Sample 2	10 bands	Sample 2
88Mdl_Sample 3	10 bands	Sample 3
88Mdl_Sample 4	10 bands	Sample 4
88Mdl_SOM_Sample 1	10 bands; SOM Class	Sample 1
88Mdl_SOM_Sample 2	10 bands; SOM Class	Sample 2
88Mdl_SOM_Sample 3	10 bands; SOM Class	Sample 3
88Mdl_SOM_Sample 4	10 bands; SOM Class	Sample 4
88Mdl_5Index_Sample 1	10 bands; 5 Derived Indices	Sample 1
88Mdl_5Index_Sample 2	10 bands; 5 Derived Indices	Sample 2
88Mdl_5Index_Sample 3	10 bands; 5 Derived Indices	Sample 3
88Mdl_5Index_Sample 4	10 bands; 5 Derived Indices	Sample 4

2.6.2. Hyper-Parameters Optimization

Hyper-parameters are the parameters that control the model training process, and they are parameters that can be optimized. Different combinations of hyper-parameters can lead to different model performance. Hyper-parameter optimization refers to the process used to find a set of hyper-parameters that yields an optimal model which minimizes the loss function on the given validation dataset. Grid search, random search, and Bayesian optimization are three of the most common approaches to perform hyper-parameter optimization. Grid search exhaustively searches all the combinations from a manually defined hyper-parameter space of a model and is the most expensive approach especially when feature space is large. Random search is a slight modification of grid search where hyper-parameter combinations are randomly selected from a manually defined hyper-parameter space instead of exhaustively enumerating of all combinations. This search method is faster than grid search, but both of them are limited to prior knowledge of hyper-parameter distribution specifications. Bayesian optimization, on the other hand, attempts to find the global optimum in minimum steps without the need to manually define each hyper-parameter sample points. Bayesian optimization builds a probabilistic model on the hyper-parameter values and the objective function, the surrogate model. A Gaussian process (GP)-based surrogate model is a popular one for Bayesian optimization because it is cheap to evaluate. The function to propose hyper-parameters combinations is referred to as the acquisition function, which is an important part of a surrogate model. The Bayesian optimization algorithm can be defined by Equation (6).

$$\begin{aligned} X_t &= argmax_X u(X|D_{1:t-1}) \\ D_{1:t-1} &= (X_1, y_1), \ldots, (X_{t-1}, y_{t-1}) \end{aligned} \tag{6}$$

where u is the acquisition function, X_t is the hyper-parameter sampling point at iteration t, $D_{1:t-1}$ is the objective function values and hyper-parameters samples from previous $t-1$ iterations. Bayesian optimization tries to find the hyper-parameter combinations at step t that is able to maximize the acquisition function u. Then the evaluation results from the objective function at step t which will be added to previous results to update the GP. Hyper-parameters for the tree-based ensemble models are optimized by Bayesian optimization. Tuned parameters includes the number of learners, the ensemble approach, the learning rate, the minimum leaf number, split criteria, the number of variables used in each node, and evaluation times.

2.7. Cloud Mask Prediction and Accuracy Evaluation

There are 4 training subsets sampled from the 88 images in the intra-group. Based on Table 2, each training subset accounts for only a very small portion of the total pixels present in the 88 images. The portion is not direly set but determined by the bin number n and the sample number m in the representative sampling stage. The portions range from 0.01% to 1.30% as shown in Table 2. Once the training processes are complete, these models are applied to the 88 images in the intra-group and the 8 images in the ultra-group to generate the predicted cloud masks. Fmask 4.0 is also applied to the 88 images in the intra-group and the 8 images in the ultra-group. The masks generated using the Fmask 4.0 algorithm include five classes which are clear land, clear water, cloud shadow, snow, and cloud. These classes are grouped into cloud, clear, and shadow to match the output of the machine learning models.

Each of the LC8 Biome images has an associated manually generated mask created by an analyst for validation purposes [6]. Previous cloud masking studies reveal that the differences due to the analyst's interpretation is about 7% [31]. To avoid this difference, all these cloud masks are digitized by a single analyst. Then, image-based confusion matrices are generated for each model and Fmask. For each image, there will be 13 confusion matrices generated, 12 of which are for the 12 model predictions, and the thirteenth for Fmask. These confusion matrices furnish the foundation of the accuracy measure calculations discussed in the next section.

3. Results

3.1. Accuracy Measurements Establish

Confusion matrices of the entire intra-group and ultra-group are generated for each of the machine learning models and the Fmask algorithm. These confusion matrices are then used to calculate the performance metrics including the correctness, the cloud commission error, and the cloud omission error at the image level and the group level. These measurements are defined in Equations (7)–(9):

$$Correctness = \frac{Cloud\ as\ Cloud + Clear\ as\ Clear + Shadow\ as\ Shadow}{Total\ pixel} \tag{7}$$

$$Cloud\ Commission\ Error = \frac{Clear\ as\ Cloud + Shadow\ as\ Cloud}{Total\ Clear + Total\ Shadow} \tag{8}$$

$$Cloud\ Omission\ Error = \frac{Total\ Cloud - Cloud\ as\ Cloud}{Total\ Cloud} \tag{9}$$

where correctness measures the overall model correctness for all three classes, cloud commission error measures the portion of pixels that are estimated as cloud but are actually not, cloud omission error measures the portion of cloud pixels that failed to be detected. Because those non-cloud pixels that are labeled as cloud will be removed in most remote sensing image analysis, cloud commission error could cause information loss. Compared to information lost, a failure of detecting cloud pixels will have a more severe negative impact on the image analysis. Therefore, controlling cloud omission error is the first priority of this research. Among all of the 88 intra-group images, only 32 images have a shadow class labeled. Limited by the training and validation size, shadow detection is not the goal of this research. Therefore, the commission error and omission error were built for cloud pixel identification.

3.2. Visual Comparison and Accuracy Assessment

For each of the 88 images we produced 14 predicted cloud masks, 12 of which are from our 12 machine learning models, one from Fmask, and one from the ground truth mask. Figure 5 illustrates four sample images with their associated ground truth image, ground truth mask, model prediction mask from model 88Mdl_5index_sample4, and Fmask. For these four sample images, both from the machine learning models and Fmask, we see that they perform well on cloud detection. However, higher commission and omission errors can be observed compared to the machine learning model prediction mask. For the forest image in the first row in Figure 5, both the prediction mask and Fmask show some degree of omission error, but Fmask has more missed cloud pixels. For the shrub image in the second row, Fmask mis-classified a large water zone as cloud. For the grass/crops image in the third row, Fmask missed some cloud pixels again in the thick cloud mixture zone. For the image in the fourth row, both masks provide a good fit based on a visual comparison.

To better explore the performance of machine learning model and Fmask, the accuracy assessment Table 4 is summarized for the four images in Figure 5. Both machine learning models and Fmask have good overall correctness (above 85%); however, Fmask has significant omission error in the first and third image, and has high commission error in the second image. This quantitative result agrees with the visual comparison result. The machine learning model has good consistent performance on the four sample images.

To provide a more comprehensive accuracy assessment, accuracy assessment tables are summarized for both the entire intra-group (Table 5) and the ultra-group (Table 6) for each of the machine learning models and for the Fmask algorithm. Tables 5 and 6 are sorted on cloud omission error in ascending order. From Table 5, the 88Mdl_SOM_sample4 performs the best on all measurements on the 88 intra-group. As can be seen in Table 6, the

88Mdl_5index_sample3 has the lowest cloud omission error and the model 88Mdl_sample3 has the best correctness performance in the ultra-group. Fmask 4.0 has the second lowest cloud omission error.

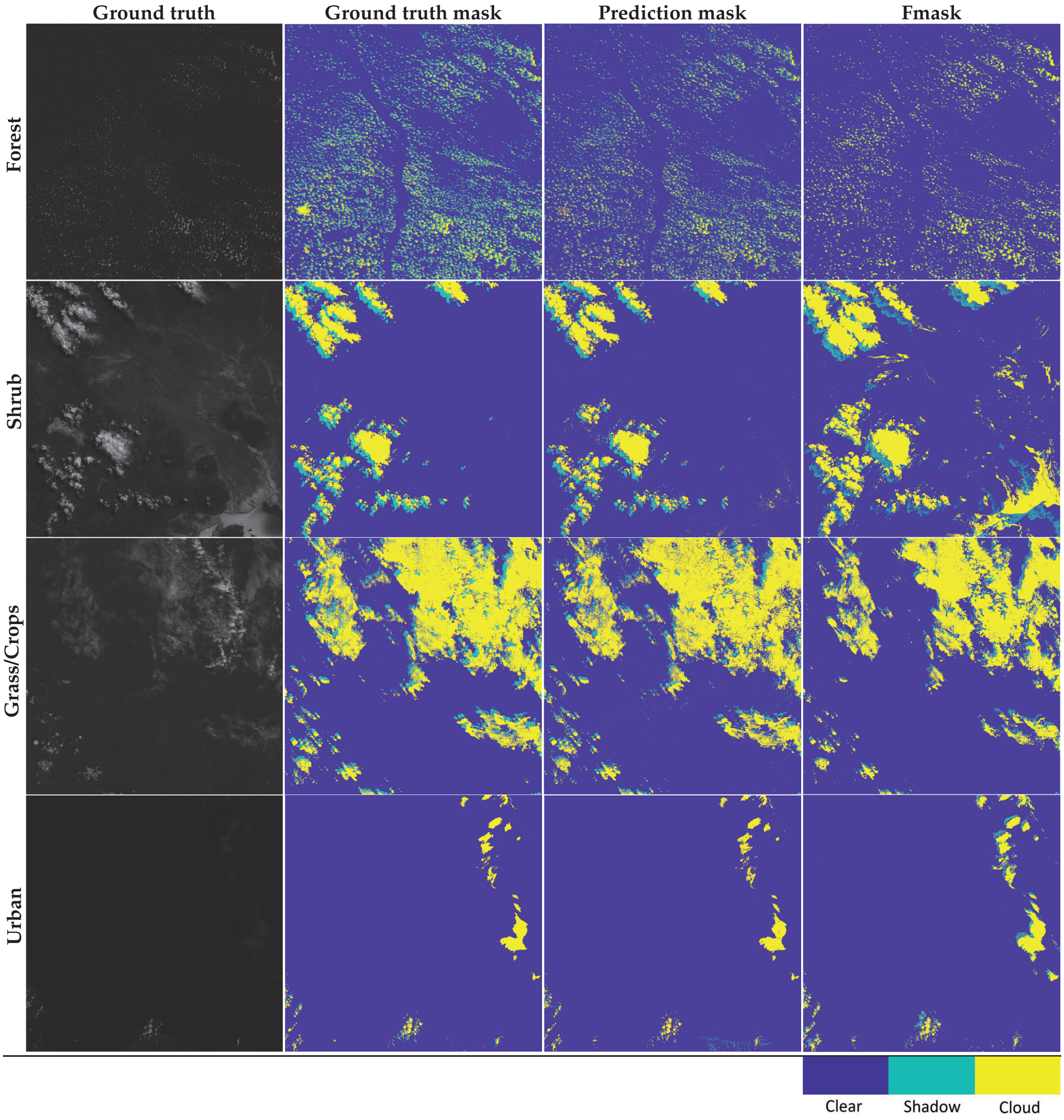


Figure 5. Sample image comparison between the validation mask, 88Mdl_5index_sample4 model predicted mask and Fmask. Each row displays an image with a certain type. The first column represents the ground truth image and the second column represents the ground truth mask, which is manually generated by expert. The third column represents the model prediction results and the fourth column displays the Fmask application result.

Table 4. Model predicted mask and Fmask accuracy comparison at image level for Figure 5. OE refers to omission error and CE refers to commission error.

Source Scene ID	Mask Type	Correct	Cloud OE	Cloud CE	Land Type
LC81750622013304LGN00	Predicted Fmask	90.30% 86.52%	24.50% 33.94%	0.12% 0.62%	Forest
LC80010732013109LGN00	Predicted Fmask	97.47% 85.39%	3.78% 0.26%	0.62% 8.73%	Shrub
LC80290372013257LGN00	Predicted Fmask	95.82% 86.65%	3.86% 27.78%	0.77% 0.66%	Grass/Crops
LC80640452014041LGN00	Predicted Fmask	99.57% 98.22%	5.18% 0.55%	0.11% 0.80%	Urban

Table 5. Intra-group accuracy assessment for machine learning models and the Fmask.

Model Names	Correct	Cloud OE	Cloud CE
88Mdl_SOM_sample4	98.57%	1.18%	0.93%
88Mdl_sample4	98.35%	1.19%	1.08%
88Mdl_5index_sample4	98.45%	1.33%	0.98%
88Mdl_5index_sample3	97.30%	1.37%	2.12%
88Mdl_sample3	98.00%	1.68%	1.38%
88Mdl_sample2	97.59%	2.15%	1.61%
88Mdl_5index_sample2	97.62%	2.16%	1.57%
88Mdl_5index_sample1	97.13%	2.50%	1.43%
88Mdl_sample1	96.76%	2.60%	1.65%
88Mdl_SOM_sample3	96.10%	3.76%	2.26%
88Mdl_SOM_sample2	95.06%	5.03%	2.87%
88Mdl_SOM_sample1	94.46%	5.22%	3.31%
Fmask 4.0	88.45%	7.57%	10.82%

Table 6. Ultra-group accuracy assessment for machine learning models and the Fmask.

Model Names	Correct	Cloud OE	Cloud CE
88Mdl_5index_sample3	86.35%	12.48%	7.96%
Fmask 4.0	83.09%	13.76%	10.69%
88Mdl_sample3	87.00%	14.78%	5.80%
88Mdl_sample2	86.81%	15.23%	5.23%
88Mdl_sample4	85.53%	15.60%	7.00%
88Mdl_SOM_sample1	86.15%	15.61%	5.84%
88Mdl_SOM_sample3	86.31%	15.67%	6.19%
88Mdl_5index_sample2	86.85%	15.72%	5.04%
88Mdl_SOM_sample4	86.00%	15.95%	6.38%
88Mdl_SOM_sample2	86.25%	15.99%	5.74%
88Mdl_5index_sample4	86.05%	16.68%	5.30%
88Mdl_sample1	84.77%	17.28%	5.06%
88Mdl_5index_sample1	85.61%	17.60%	3.99%

As can be seen in Table 6, Fmask has the lowest performance for all metrics when compared to the machine learning models. To have a in-depth exploration on the factors influencing the Fmask performance, the eight images with the lowest performance are summarized (Table 7). Four of the eight images are plotted in Figure 6. The accuracy measurements are summarized at image level for the the Fmask and the five indices machine learning model trained on sample4 for comparison. The machine learning model performed well on all images, while the Fmask has high variability across the images. The overall correctness could decrease to as low as 7.47% for the Ice/Snow images. Four out of the eight images are for the ice/snow land type. These images have the common features that the background reflectance is “bright”. This feature causes trouble for Fmask in two ways. In some images, Fmask mis-classifies ice/snow as cloud, while in some images Fmask mis-classifies cloud as clear pixels. This quantitative finding conforms with what one can see by visual inspection of the results in Figure 6, where we see that Fmask has difficulty in detecting cloud in images in the first and third row. In the first row, Fmask mis-classifies the cloud pixels as snow/ice-covered clear pixel. In the third row, Fmask mis-classifies snow/ice-covered clear pixels as cloud. As can be seen in the fourth row of Figure 6, water is another case with "bright" background, where Fmask mis-classifies large water regions as cloud pixels.

The other three images types listed in Table 7 are grass/crops, forest, and shrubland. The grass/crops, forest and shrub images are similar in that they absorb a large fraction of the visible bands and make the images “dark”. This effect could influence the performance of Fmask. As can be seen in the Table 7, Fmask has high cloud omission error over these three land types. The machine learning model, on the other hand, has a stable performance across the different land types. The cloud commission errors are NaN in the table because this image is 100% covered by thin cloud; thus, by definition, no cloud commission error can be calculated. For the shrubland image, Fmask mis-classifies large thin cloud zones as clear pixels (second row of Figure 6).

Table 7. Model predicted mask and Fmask accuracy comparison for low Fmask accuracy images.

Source Scene ID	Mask Type	Correct	Cloud OE	Cloud CE	Land Type
LC80060102014147LGN00	Predicted Fmask	98.35% 69.67%	0.00% 57.33%	0.00% 6.54%	Ice/Snow
LC80211222013361LGN00	Predicted Fmask	97.65% 55.15%	2.70% 73.21%	1.90% 5.99%	Ice/Snow
LC80290372013257LGN00	Predicted Fmask	95.82% 86.65%	3.86% 27.78%	0.77% 0.66%	Grass/Crops
LC81720192013331LGN00	Predicted Fmask	100.00% 64.92%	0.00% 35.08%	nan nan	Forest
LC82271192014287LGN00	Predicted Fmask	99.90% 7.47%	0.92% 26.71%	0.04% 97.04%	Ice/Snow
LC81490122013218LGN00	Predicted Fmask	99.27% 18.49%	0.35% 87.72%	5.45% 0.74%	Shrubland
LC80200462014005LGN00	Predicted Fmask	96.88% 90.03%	8.71% 15.42%	0.09% 0.81%	Ice/Snow
LC80210072014236LGN00	Predicted Fmask	98.21% 79.96%	8.35% 15.79%	0.83% 16.80%	Water

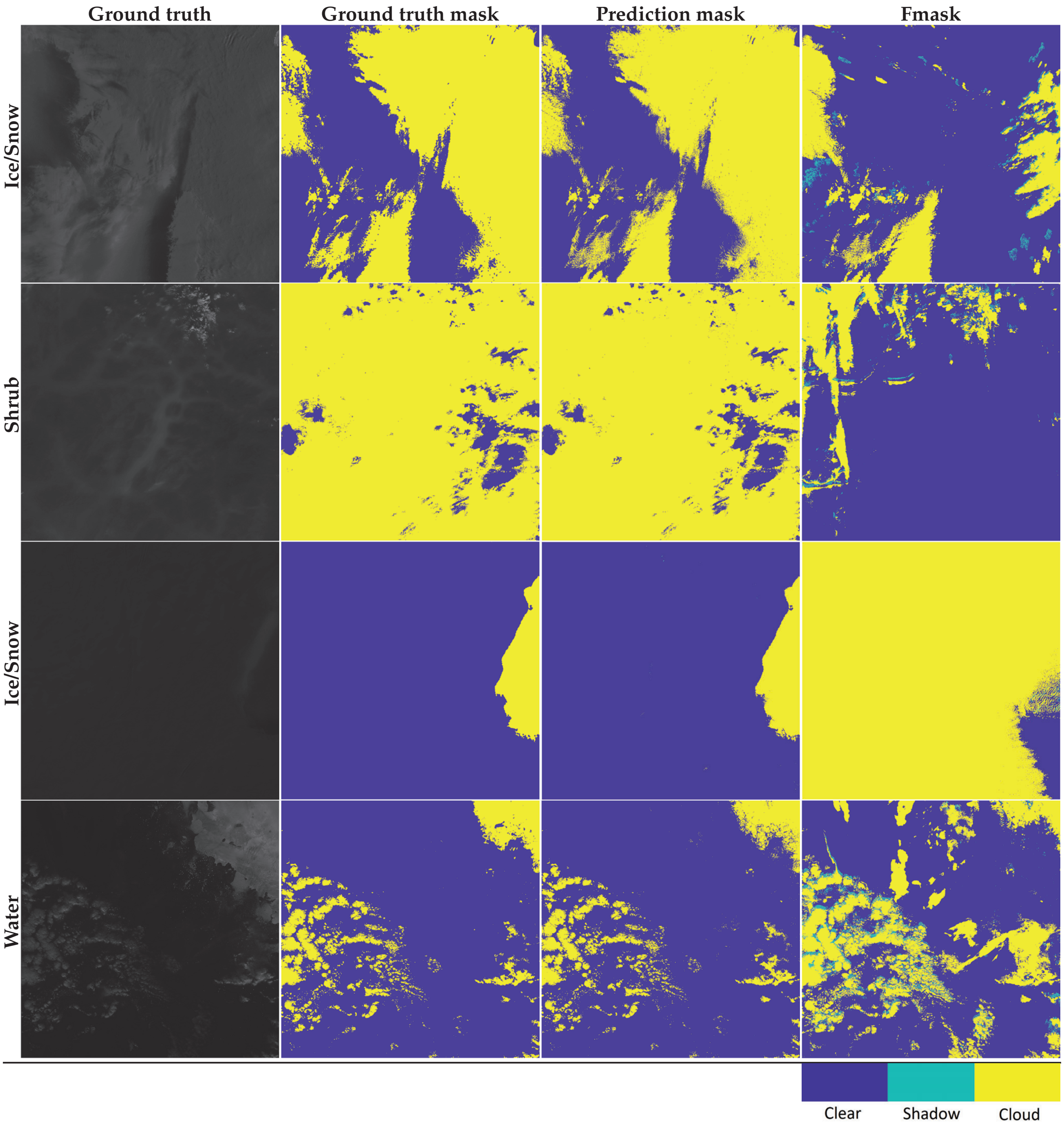
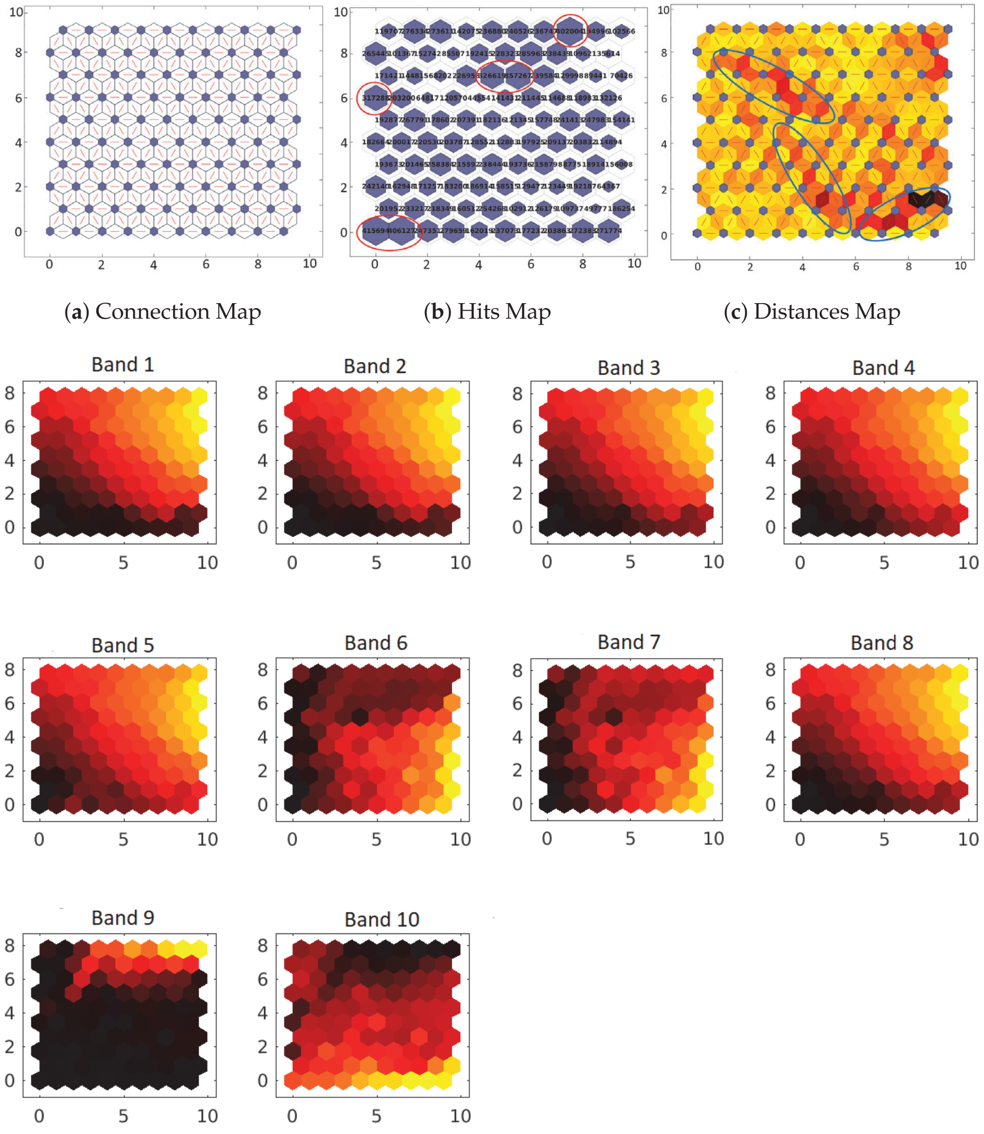


Figure 6. Fmask miss-classified images. Each row displays an image with a certain type. The first column represents the ground truth image and the second column represents the ground truth mask. The third column represents the model prediction results and the fourth column displays the Fmask application result.

3.3. Feature Importance and Topology Analysis

Machine learning has been traditionally treated as a "black box", because most of the machine learning algorithms focus on solving problems based on training data but do not readily give interpretations on how the input variables influence the prediction outcome. The unsupervised SOM classification method and the supervised tree-based ensemble classification method provides ways to explore these influences of the input variables on the prediction outcomes. As in Figure 7, a 10-by-10 network is constructed, which includes 100 neurons. There is a weight vector associate with each neuron and the weight vector has the same dimension as input vectors. As the training process goes, these weight vectors move to the center of input vector clusters. Once the training process completes, the SOM model is applied to the training dataset to explore the variable characteristics. Each neuron represents a cluster, and each input vector will be assigned to one of the 100 neurons. A useful feature of SOMs is topology preservation, which means those close vectors in the input space remain close together after being projected to 2D planes [32]. Therefore, a high dimensional vectors can be visualized in a 2D plane.



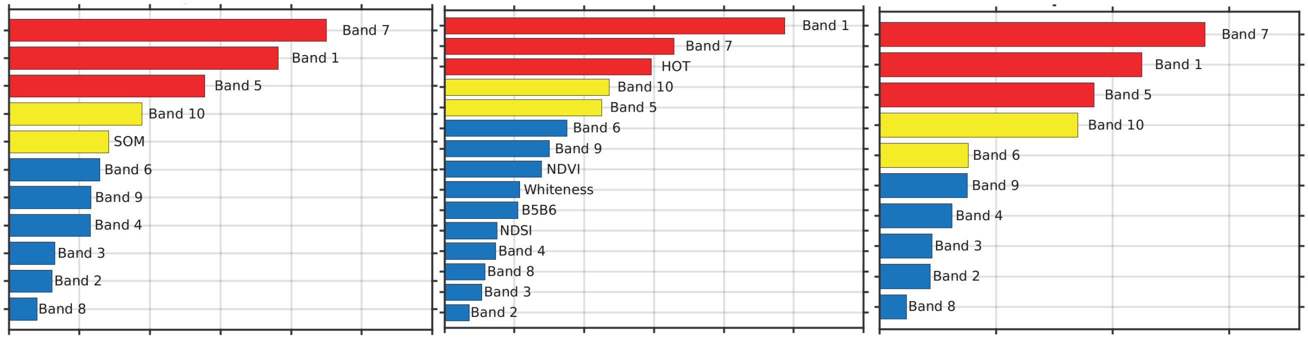
(**a**) Connection Map (**b**) Hits Map (**c**) Distances Map

(**d**) Weight Planes Map

Figure 7. The four SOM plots. A total of 100 neurons are plotted on a 10 by 10 gridded plane. Axis x and axis y are the neuron location indices. (**a**) The connection map displays the neuron topology, (**b**) the hits map represents the number of vectors associated with each neuron, (**c**) the SOM neighbor weight distance map displays the distance between each neuron, and (**d**) displays the weight of each input component (spectral bands) in a 2D plane.

Figure 7a illustrates the SOM neuron structure in a 2D gridded plane. As can be seen in Figure 7b, the SOM hits map displays how many input vectors fall within each SOM class. There are several large clusters of vectors annotated with red circles which have hit numbers of greater than 300,000. The Figure 7c indicates the feature space distance between each node. Black and dark red represent longer inter-SOM class separation distances while yellow and orange represent shorter inter-SOM class separation distances. Several red and black connections marked with blue ovals segment the whole input vectors into different zones. Each zone shares some degree of similarity with the adjacent classes. The black regions indicate the longest inter-class separations. Figure 7d displays the weights of each input spectral bands on the neuron plane. Similar patterns in weight planes indicate high correlation between the different bands. As can be seen, band 1, band 2, band 3, band 4, band 5 and band 8 indicate a strong correlation pattern in the 2D topology. Band 6 and band 7 share some degree of similarly. On the other hand, band 9 and band 10 give unique information to the SOM model. These plots give information of how the input variables influence the unsupervised classification and the similarity between classes. The SOM model is able to integrate these many input variables and provide new insights into the data, which can be used as an additional input feature for the tree-based models to enhance their performance.

One advantage of tree-based models is that they are easy to interpret. The decision tree algorithms internally select the variable which can most reduce the entropy at each decision tree splitting node. The variable that can most reduce the entropy will be put in the root node, and the second most important variable will be placed next after the root node. Thus, variables are sorted based on their capability to reduce the entropy. As a result, the importance and contribution of each variable to the model can be easily traced and explained. Split nodes are determined by their capability to reduce the entropy of the leaf node. Therefore, the importance rank of input variables in differentiating clusters can be traced and visualized. Figure 8 displays the importance rank plot for the three best performing ensemble models, which are the SOM model trained with sample 4, the five indices model trained on sample 3, and the base model with only 10 wavelength bands trained on sample 3. Among these tree models, band 7, band 1, band 5 and HOT have the most influence on the predicted model class. For the model incorporating the SOM classification, the SOM class is the fifth most important input feature.



(**a**) Plot for 88Mdl_SOM_sample4 (**b**) Plot for 88Mdl_5index_sample3 (**c**) Plot for 88Mdl_sample3

Figure 8. The importance plot for three best models, which are (**a**) the model trained on sample 4 with SOM as input, (**b**) the model with 5 indices trained on sample 3 and (**c**) the base model with 10 wavelength bands trained on sample 3. Model (**a**) has the best performance on all metrics in the intra-group, while model (**b**,**c**) has the best performance on omission error and correctness, respectively. Red and yellow bars represent the top 5 important features and blue bars represent the importance of the rest variables.

3.4. Algorithm Complexity and Running Time Analysis

The proposed ensemble method is based on a decision tree model; thus, the algorithm complexity of decision tree is the conceptual base for the ensemble algorithm complexity analysis. The depth complexity of a fully developed tree would be in $O(log(n))$ for a balanced tree, and $O(n)$ for a extremely unbalanced tree, where n is the number of training samples. In practice, decision trees are rarely fully developed or extremely unbalanced. As a result, $O(log(n))$ is a good approximation for the tree depth complexity. A quality function is calculated for every node for each feature at all levels. The algorithm complexity for a decision tree would be $O(m \times n \times log(n))$, where m is the number of features. For bagging ensemble methods, random forest techniques are applied, where only the square root number of features are considered on each node for each tree. Thus, the training complexity is $O(k \times m^{\frac{1}{2}} \times n \times log(n))$ where k is the number of trees in the ensemble method. For Boosting ensemble methods, all the features are considered for model training, thus the complexity is $O(k \times m \times n \times log(n))$. Selection of the Boosting or Bagging approach are two of the ensemble hyperparameters which are determined by the hyperparameter optimization. Once the the model is trained and ready for running, the run time complexity for ensemble types are $O(k \times m)$.

The running time is summarized in Table 8. Training time is the total elapsed time for all the steps to generate a model including representative training subsets sampling, model training, hyper-parameter optimization, plotting and model saving under Intel(R) Xeon(R) CPU E7-4850 v3 @ 2.20 GHz. The Running time is the average elapsed time to make a cloud mask prediction and to plot the prediction masks under Intel(R) Xeon(R) CPU X5650 @ 2.67 GHz.

Table 8. The model training and running time.

Model Name	Training Time (s)	Running Time (s)
88Mdl_Sample1	971	836
88Mdl_Sample2	3037	133
88Mdl_Sample3	42,279	346
88Mdl_Sample4	694,051	1037
88Mdl_SOM_sample1	1587	268
88Mdl_SOM_sample2	5099	268
88Mdl_SOM_sample3	33,839	303
88Mdl_SOM_sample4	524,561	371
88Mdl_5Index_sample1	2162	326
88Mdl_5Index_sample2	4348	251
88Mdl_5Index_sample3	20,334	547
88Mdl_5Index_sample4	355,966	167

3.5. Hyperparameter Sensitivity Analysis

Hyperparameters are important parameters that control the behavior of empirical machine learning models. The hyperparameter optimization seeks to suppress the variance and bias of a model by finding a set of global optimal hyperparameters that minimize the objective function value. We proposed the Bayesian hyperparameters optimization which minimizes the cross-validation classification error with 30 iterations for each model. Optimized hyperparameters as well as the five-fold cross-validation objective function values measuring the model errors are summarized for each model. Hyperparameters at the initial iteration and at the best iteration are listed in the Table 9 to demonstrate how the optimization process improves the objective function value. Each model has two iterations result listed. The first iteration with the value one is the initial round of hyperparameter optimization, while the second number refers to the iteration where the best optimization results are achieved among the total 30 iterations. Ensemble column refers to the ensemble method including Bagging and Boosting two categories. Trees represents the number of single decision trees included in each model. LearnRate is a hyperparameter controlling the learning speed for the Boosting method only. MinLeafSize is the minimum number of samples required to be at a leaf node. The MaxNumSplits is the maximal number of decision splits for each tree. The MinLeafSize and MaxNumSplits together control the depth of tree models. As in Table 9, the objective function values decreased significantly after hyperparameter optimization for all the 12 ensemble models. Compared to the initial model, the optimized models are dominated by employing the Boosting method, with larger numbers for the MinLeafSize and more MaxNumSplits.

Table 9. Hyperparameters sensitivity analysis.

Model Name	Iteration	Objective	Ensemble	Trees	LearnRate	MinLeafSize	MaxNumSplits
88Mdl_Sample1	1 25	0.99 0.03	RUSBoost Bag	91 31	0.07 NA	8435 1	1741 124,550
88Mdl_Sample2	1 28	0.99 0.02	RUSBoost AdaBoostM2	25 25	0.00 0.40	153,080 39	19 3153
88Mdl_Sample3	1 11	0.99 0.02	RUSBoost RUSBoost	13 12	0.02 0.02	51,723 7	61,518 647,140
88Mdl_Sample4	1 19	0.13 0.02	RUSBoost RUSBoost	17 145	0.05 0.02	5433 1	37,578 1.77×10^7
88Mdl_SOM_sample1	1 27	0.23 0.03	RUSBoost Bag	12 15	0.71 NA	13 10	2 177,740
88Mdl_SOM_sample2	1 24	0.16 0.03	RUSBoost AdaBoostM2	13 27	0.02 0.00	4 11	9 7937
88Mdl_SOM_sample3	1 21	0.06 0.02	RUSBoost RUSBoost	19 218	0.01 0.05	4 1	446,920 28,364
88Mdl_SOM_sample4	1 23	0.09 0.02	RUSBoost RUSBoost	23 88	0.01 0.13	369 1	3.07×10^6 1.32×10^7
88Mdl_5Index_sample1	1 26	0.22 0.03	Bag Bag	14 29	NA NA	62,299 22	65,065 2138
88Mdl_5Index_sample2	1 26	0.99 0.02	RUSBoost AdaBoostM2	28 12	0.04 0.80	542,860 16	95 418,850
88Mdl_5Index_sample3	1 30	0.99 0.02	RUSBoost RUSBoost	28 412	0.04 0.11	542,860 2	95 255,040
88Mdl_5Index_sample4	1 16	0.99 0.01	RUSBoost RUSBoost	18 276	0.27 0.60	1.78×10^6 7	2.89×10^6 420,910

4. Discussion

This paper established and evaluated 12 ensemble machine learning models for pixel-based cloud classification trained and tested using the labeled Landsat 8 Biome dataset. In addition to the information from the ten Landsat bands some additional input features were used, including an unsupervised self-organizing map classification and five indices derived from various band combinations. This feature engineering improved the performance of the machine learning cloud pixel classification.

Typically, training a machine learning model by using as much data as possible enables a better model performance. However, when it comes to supervised cloud classification, we are often constrained by the availability of labeled images and sometimes by the available computing power. We therefore designed a strategy to use only a tiny subset of the total number of pixels available for model training.

The objective of this research is three-fold. First, we described, implemented and validated an ensemble approach to build supervised classification model for cloud detection, including model selection, selection of representative training data, and feature engineering. Second, we explored the importance of the input features. Last but not least, we investigated how the size of the training subsets influenced the model performance.

Generally, a larger number of training samples enables better model performance when using the intra-group, while the training size has less influence on the model performance of the ultra-group. In addition, our ensemble models had consistent performance on all land types in contrast to Fmask which had trouble in differentiating cloud pixels in some "bright" and "dark" area such as snow/ice, shrub, and some forest area. The fact that the overall model performances on the intra-group are better than the performance on the ultra-group is reasonable because all the training data are sampled from the intra-group, even though just a small fraction of the total available pixels were used. As shown in the Tables 2 and 5, models trained on subsets that only accounted for 1.3% of the total pixels available could achieve an accuracy as high as 98.57% when applied to the whole 88 intra-group images. For the independent validation dataset that our machine learning models had never previously seen, even though the performance decreased, they still out-performed Fmask.

Landsat 8 images could suffer from various degradation, noise, or variability during the image processing. The generation capability of the machine learning models have been largely determined by the representatives of the training samples. In this study, the LC8Biome cloud validation data are designed to cover the geographical scope of the world and the eight biome types, and includes three levels of cloud coverage (clear, mid, cloudy). This approach makes the training set representative enough to allow a model with good generalization for the normal radiance variability of LC8 data. However, the LC8 Biome dataset is free from severe noise, which is caused by such as equipment failure, which makes the cloud detection on images with severe noise out of the scope of the proposed machine learning models. Nevertheless, the only well-known issue so far for LC8 is the thermal band failure. Although the lack of cloud validation masks for severely distorted image stops us from doing an accuracy assessment; the incorporation of 10 wavelength bands as predictors in the machine learning models allows some degree of resistance to certain band failure. In addition, the proposed machine learning method is also suitable to establish a cloud detection model for images with high noise once the training samples with severe noise are available.

5. Conclusions

Overall, the tree-based ensemble machine learning models with Bayesian optimization achieved better performance than Fmask for all metrics over the 88 images in the intra-group. On the eight images in the ultra-group, 88Mdl_5index_sample3 has the lowest cloud omission error and the 88Mdl_sample3 has the best correctness. The base model with 10 wavelength bands as the input variable could achieve good performance for both the intra-group and ultra-group. Fmask had issues in differentiating cloud pixels mainly in ice/snow images, and some in the water, forest, and shrub images.

The empirical pixel-based machine learning models discussed in this paper could achieve very good and consistent performance when using only a tiny portion of the total available training data. This result indicates that a pixel-based method can perform well, and that each scene obviously has unique spectral characteristics, and having a small portion of example pixels from each of the sub-regions in a scene can improve the model accuracy significantly. Integrating the self-organizing map results and five band derived indices as a part of input variables could further help to reduce the cloud commission error and cloud omission error compared to the base models. Based on hyperparameter sensitivity analysis, the Boosting ensemble method with a small MinLeafSize and large MaxNumSplits are effective settings for high accuracy model training. Among all the input variables investigated in this research, band 1, band 5, band 7, and HOT are the five most important variables for the best three models; Band 10, band 6, and SOM are in the second tier; band 4, band 6, band 9, $NDVI$, whitness, $B5/B6$ ratio are in the third tier which also contributes to the model performance.

Author Contributions: Conceptualization, X.Y. and D.J.L.; methodology, X.Y. and D.J.L.; software, X.Y. and D.J.L.; validation, X.Y.; formal analysis, X.Y.; investigation, X.Y.; resources, X.Y. and D.J.L.; data curation, X.Y.; writing—original draft preparation, X.Y.; writing—review and editing, D.J.L.; visualization, X.Y.; supervision, D.J.L.; project administration, X.Y. and D.J.L.; funding acquisition, D.J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The LC8-Biome cloud validation dataset is available at URL: https://landsat.usgs.gov/landsat-8-cloud-cover-assessment-validation-data (accessed on 18 August 2021). The code for image processing, model training and model application is available at URL: https://github.com/xiaoheyu (accessed on 18 August 2021).

Acknowledgments: The concept for this paper was first developed as an example of the use of machine learning for the ESA ϕ-lab machine learning bootcamp for remote sensing in January 2019. URL: https://philab.phi.esa.int/ (accessed on 18 August 2021). Christopher Simmons is gratefully acknowledged for his computational support. 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 on 18 August 2021). The authors acknowledge the Texas Advanced Computing Center (TACC) at the University of Texas at Austin for providing HPC, visualization, database, or grid resources that have contributed to the research results reported within this paper. URL: http://www.tacc.utexas.edu (accessed on 18 August 2021).

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

References

1. King, M.D.; Platnick, S.; Menzel, W.P.; Ackerman, S.A.; Hubanks, P.A. Spatial and Temporal Distribution of Clouds Observed by MODIS Onboard the Terra and Aqua Satellites. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 3826–3852. [CrossRef]
2. Arvidson, T.; Goward, S.; Gasch, J.; Williams, D. Landsat-7 long-term acquisition plan. *Photogramm. Eng. Remote Sens.* **2006**, *72*, 1137–1146. [CrossRef]
3. Irish, R.R. Landsat 7 automatic cloud cover assessment. Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imagery VI. *Int. Soc. Opt. Photonics* **2000**, *4049*, 348–355.
4. Zhu, Z.; Woodcock, C.E. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sens. Environ.* **2012**, *118*, 83–94. [CrossRef]
5. Zhu, Z.; Wang, S.; Woodcock, C.E. Improvement and expansion of the Fmask algorithm: Cloud, cloud shadow, and snow detection for Landsats 4–7. 8, and Sentinel 2 images. *Remote Sens. Environ.* **2015**, *159*, 269–277. [CrossRef]
6. Foga, S.; Scaramuzza, P.L.; Guo, S.; Zhu, Z.; Dilley, R.D., Jr.; Beckmann, T.; Schmidt, G.L.; Dwyer, J.L.; Hughes, M.J.; Laue, B. Cloud detection algorithm comparison and validation for operational Landsat data products. *Remote Sens. Environ.* **2017**, *194*, 379–390. [CrossRef]
7. Hughes, M.J.; Hayes, D.J. Automated detection of cloud and cloud shadow in single-date Landsat imagery using neural networks and spatial post-processing. *Remote Sens.* **2014**, *6*, 4907–4926. [CrossRef]
8. Li, Z.; Shen, H.; Li, H.; Xia, G.; Gamba, P.; Zhang, L. Multi-feature combined cloud and cloud shadow detection in GaoFen-1 wide field of view imagery. *Remote Sens. Environ.* **2017**, *191*, 342–358. [CrossRef]
9. Zhu, Z.; Woodcock, C.E. Automated cloud, cloud shadow, and snow detection in multitemporal Landsat data: An algorithm designed specifically for monitoring land cover change. *Remote Sens. Environ.* **2014**, *152*, 217–234. [CrossRef]
10. Candra, D.S.; Phinn, S.; Scarth, P. Automated Cloud and Cloud-Shadow Masking for Landsat 8 Using Multitemporal Images in a Variety of Environments. *Remote Sens.* **2019**, *11*, 2060. [CrossRef]
11. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
12. Xie, F.; Shi, M.; Shi, Z.; Yin, J.; Zhao, D. Multilevel cloud detection in remote sensing images based on deep learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 3631–3640. [CrossRef]
13. An, Z.; Shi, Z. Scene learning for cloud detection on remote-sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 4206–4222. [CrossRef]
14. Zhang, Q.; Xiao, C. Cloud detection of RGB color aerial photographs by progressive refinement scheme. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 7264–7275. [CrossRef]

15. Li, Z.; Shen, H.; Wei, Y.; Cheng, Q.; Yuan, Q. Cloud detection by fusing multi-scale convolutional features. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *4*, 149–152. [CrossRef]
16. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
17. Segal-Rozenhaimer, M.; Li, A.; Das, K.; Chirayath, V. Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (CNN). *Remote Sens. Environ.* **2020**, *237*, 111446. [CrossRef]
18. Zi, Y.; Xie, F.; Jiang, Z. A cloud detection method for Landsat 8 images based on PCANet. *Remote Sens.* **2018**, *10*, 877. [CrossRef]
19. Yang, J.; Guo, J.; Yue, H.; Liu, Z.; Hu, H.; Li, K. CDnet: CNN-based cloud detection for remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6195–6211. [CrossRef]
20. Guo, J.; Yang, J.; Yue, H.; Tan, H.; Hou, C.; Li, K. CDNetv2: CNN-Based cloud detection for remote sensing imagery with cloud-snow coexistence. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 700–713. [CrossRef]
21. Shahraki, F.F.; Prasad, S. Graph convolutional neural networks for hyperspectral data classification. In Proceedings of the 2018 IEEE global conference on signal and information processing (GlobalSIP), Anaheim, CA, USA, 26–29 November 2018; pp. 968–972.
22. Qin, A.; Shang, Z.; Tian, J.; Wang, Y.; Zhang, T.; Tang, Y.Y. Spectral–spatial graph convolutional networks for semisupervised hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 241–245. [CrossRef]
23. Pu, S.; Wu, Y.; Sun, X.; Sun, X. Hyperspectral Image Classification with Localized Graph Convolutional Filtering. *Remote Sens.* **2021**, *13*, 526. [CrossRef]
24. Hong, D.; Gao, L.; Yao, J.; Zhang, B.; Plaza, A.; Chanussot, J. Graph convolutional networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 5966–5978. [CrossRef]
25. Wang, B.; Ono, A.; Muramatsu, K.; Fujiwara, N. Automated detection and removal of clouds and their shadows from Landsat TM images. *IEICE Trans. Inf. Syst.* **1999**, *82*, 453–460.
26. Mahajan, S.; Fataniya, B. Cloud detection methodologies: Variants and development—A review. *Complex Intell. Syst.* **2020**, *6*, 251–261. [CrossRef]
27. U.S. Geological Survey. *L8 Biome Cloud Validation Masks*; U.S. Geological Survey: Reston, VA, USA, 2016. [CrossRef]
28. Gómez-Chova, L.; Camps-Valls, G.; Calpe-Maravilla, J.; Guanter, L.; Moreno, J. Cloud-screening algorithm for ENVISAT/MERIS multispectral images. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 4105–4118. [CrossRef]
29. Zhang, Y.; Guindon, B.; Cihlar, J. An image transform to characterize and compensate for spatial variations in thin cloud contamination of Landsat images. *Remote Sens. Environ.* **2002**, *82*, 173–187. [CrossRef]
30. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **1982**, *43*, 59–69. [CrossRef]
31. Scaramuzza, P.L.; Bouchard, M.A.; Dwyer, J.L. Development of the Landsat data continuity mission cloud-cover assessment algorithms. *IEEE Trans. Geosci. Remote Sens.* **2011**, *50*, 1140–1154. [CrossRef]
32. Kiviluoto, K. Topology preservation in self-organizing maps. In Proceedings of the International Conference on Neural Networks (ICNN'96), Washington, DC, USA, 3–6 June 1996; Volume 1, pp. 294–299.

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 $\approx 2\ \mu m$). 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 µm). For each of these six parameters we created an empirical multivariate non-linear non-parametric machine learning regression model with hyper-parameter optimization.

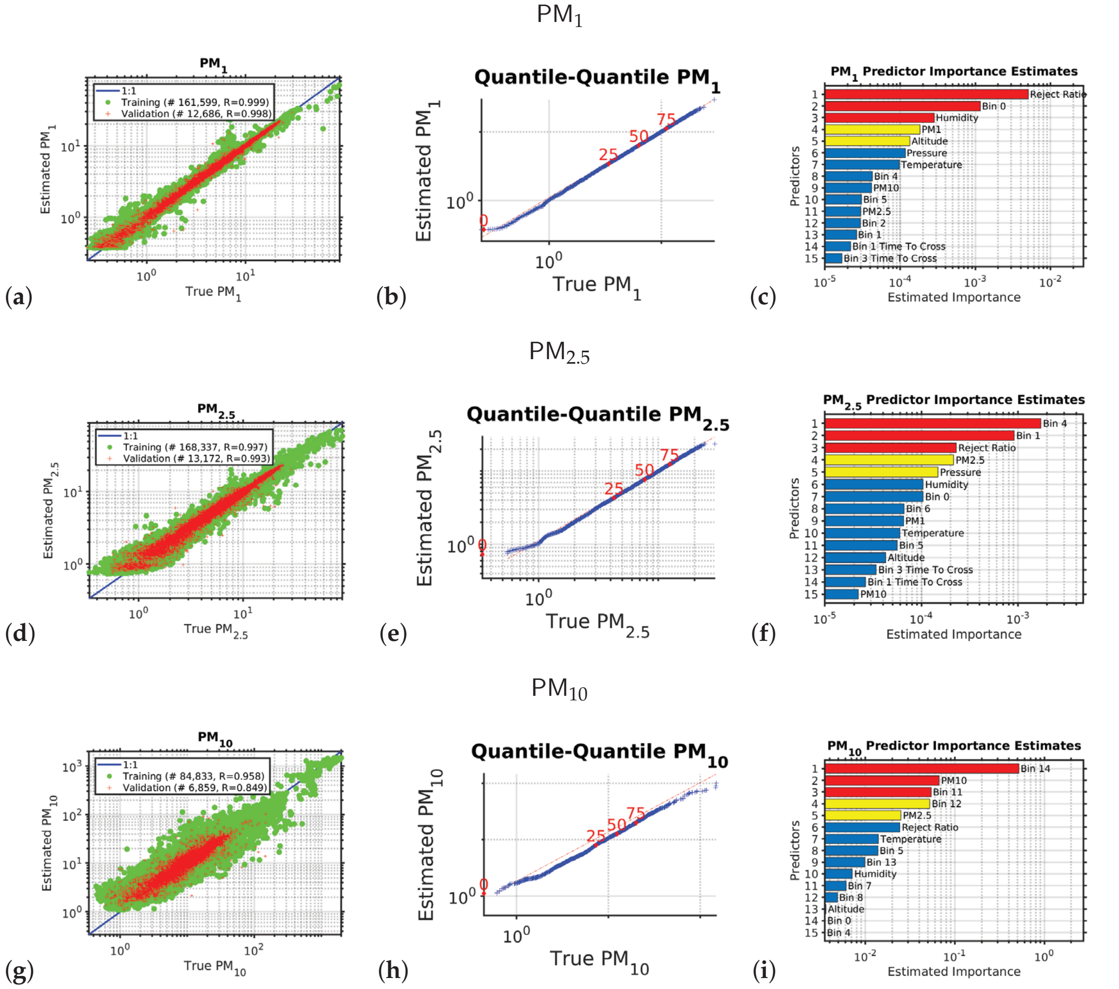


Figure 1. This figure shows the results of the multivariate non-linear non-parametric machine learning regression for particulate matter PM_1 (panels (**a**)–(**c**)), $PM_{2.5}$ (panels (**d**)–(**f**)), and PM_{10} (panels (**g**)–(**i**)). The left hand column of plots shows the log–log axis scatter 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.

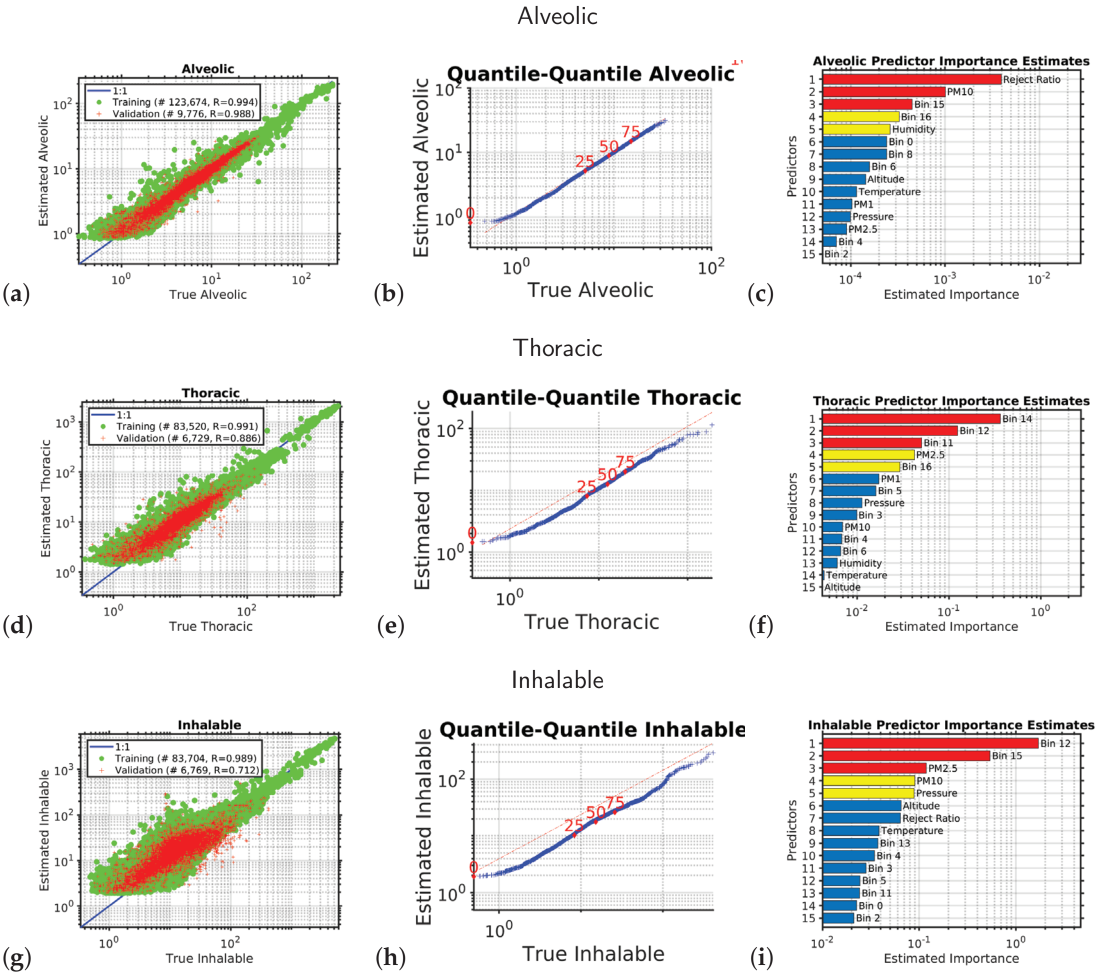


Figure 2. This figure shows the results of the multivariate non-linear non-parametric machine learning regression for the alveolic (panels (**a**)–(**c**)), thoracic (panels (**d**)–(**f**)), and inhalable size fractions (panels (**g**–**i**)). The left hand column of plots shows the log–log axis scatter 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]

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).