Pacific Northwest
NATIONAL LABORATORY

PNNL-32673

Machine learning approaches to streamline and enhance the analysis of multiscale imaging data for bioaerosol and soil particles

March 2022

Tamas Varga
Sean M. Colby
Swarup China
Anil K. Battu

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical
Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-32673

Machine learning approaches to streamline and enhance the analysis of multiscale imaging data for bioaerosol and soil particles

March 2022

Tamas Varga
Sean M. Colby
Swarup China
Anil K. Battu

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

Bioaerosol and soil particles are ubiquitous in the environment. They are multicomponent and complex in nature displaying mixed inorganic and organic components. The way components are mixed in a bioaerosol sample is referred to as its mixing state. Soil particles are also a mixture of inorganic (mineral) and organic (soil organic matter) components. Bioaerosol particles contribute to a major fraction of coarse mode atmospheric particles, especially in the tropical areas, contributing up to 80 % of the particle mass concentration. The mixing state of particles is crucial to evaluate because it impacts several important environmental processes such as warm and cold cloud formation and radiation budget. Mixing states in aerosols are accompanied by chemical reactions across solid-liquid-gas interfaces. In this study, we utilized elemental compositions and microcopy images of thousands of atmospheric particles acquired by computer-controlled scanning electron microscope equipped with an energy-dispersive x-ray spectrometer to compute the mixing state of atmospheric particles. A 2D convolutional neural network (CNN), also known as convnet, was used to model the relationship between low resolution imaging data and higher resolution spectroscopy data, with the former as training input and the latter as target output. Two types of CNNs were implemented and tested; a basic CNN and an Inception-v3 network. For binary classification, the basic CNN achieved an accuracy of 84.29 % across all atom types, and the Inception-v3-like network achieved an accuracy 85.51 %. This study demonstrates the applicability of deep learning to handle large amounts of imaging/chemical spectroscopy data efficiently and evaluate particle mixing state from a range of environmental samples.

Summary

A 2D convolutional neural network (CNN), also known as convnet, was used to model the relationship between low resolution imaging data and higher resolution spectroscopy data, with the former as training input and the latter as target output. Two types of CNNs were implemented and tested; a basic CNN and an Inception-v3 network. For binary classification, the basic CNN achieved an accuracy of 84.29 % across all atom types, and the Inception-v3-like network achieved an accuracy 85.51 %. The above-described platform will enable the efficient, streamlined analysis of thousands of particles by reducing analysis time, operator bias and error, and being more cost-effective. Next steps in determining/predicting particle mixing states are planned to: (1) enhance deep learning with data input from x-ray imaging and x-ray absorption spectroscopy, (2) experimentally verify predictions, and (3) introduce a multiscale aspect to this work. The latter involves the deep learning challenge of modeling the relationship between low-resolution image data and higher-resolution chemical information from spectroscopy. Extending the input images to 3D data for component segmentation within particles/aggregates would also be a valuable endeavor. As far as applications of this platform goes, we believe our network can be used for different data/ material systems as well as different instrumentation with small modifications to help users process large data sets.

Acknowledgments

The research was conducted under the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory (PNNL), a multi-program national laboratory operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RL01830. Part of the work was performed at EMSL, a DOE Office of Science User Facility sponsored by the Office of Biological and Environmental Research.

Acronyms and Abbreviations

2D: two-dimensional

Al: Aluminum

C: Carbon

CNN: convolutional neural network

Fe: Iron

N: Nitrogen

O: Oxygen

P: Phosphorus

Si: Silicon

Contents

Figures

1.0 Introduction

One of the challenges in data analytics related to the biological and environmental sciences is our limited capacity to process large amounts of combined imaging and chemical data collected on different samples, often at different length scales. One common example is when microscopic imaging is combined with some form of elemental analysis, where optical or x-ray absorption-based density needs to be correlated with elemental maps and corresponding spectroscopic information (**Fig. 1**). So far, data processing and analysis has mostly been done by heavy, time-consuming involvement of the operator using classical image processing, and spectral analysis techniques. There is a need for more **efficient handling of large amounts of related imaging/chemical spectroscopy data**, and for the **capability to predict chemical information from images using data analytics**.

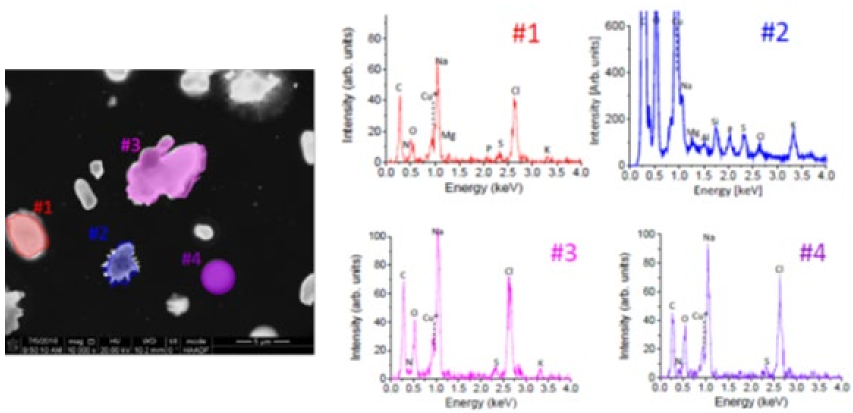

Figure 1. Example of correlated microscopic images and spectroscopy data for atmospheric aerosol particles

Aerosolized biological particles (bioaerosol) and soil particles are ubiquitous in the environment. Bioaerosol particles originate from the biosphere (pollen, bacteria, fungal spores, fragments of living organisms, soil, etc.), and they significantly influence the biosphere, the atmosphere, and public health [1-4]. They contribute to a major fraction of coarse (2-3 µm size) atmospheric particles, especially in the tropical areas, contributing up to 80 % of the particle mass concentration [2, 5-6]. By their impact on cloud and ice formation [1, 6-8] biological particles also influence the Earth’s energy budget by absorbing and scattering radiation from the Sun [1, 9]. Soil, also commonly referred to as earth or dirt, is a mixture of organic matter, minerals, gases, liquids, and organisms that together support life. Earth's body of soil serves as a medium for plant growth, a means for water storage, supply, and purification, a modifier of the atmosphere (connection to bioaerosols) as well as a habitat for organisms [10]. A significant portion of atmospheric particles are dust from soil.

Bioaerosol and soil particles are multicomponent and complex in nature; they display mixed inorganic (mineral) and organic components. The way components are mixed in an aerosol sample is called its mixing state. Soils are also a mixture of inorganic (mineral) and organic (soil organic matter) components. In aerosol particles, the mixing state is crucial to evaluate because it impacts several important environmental processes such as warm and cold cloud formation and radiation budget. We focused on bioaerosol particles from the Amazon rainforest. The objective of this project was to develop a data analysis platform for analysis of these complex bioaerosol particles that would allow for the analysis of mixing states, identification of different classes of fungal spores and bacteria, as well and image-based predictions on particle chemistry. The specific problems we wanted to solve were: *(1) Increase our ability to down select useful information/region of interest from lower resolution images of bioaerosol or soil particles for subsequent spectroscopic analysis at higher resolution; (2) Identify particle classes (e.g. different classes of spores and bacteria) and evaluate mixing state (mixture of organic and inorganic particles) and chemical associations within particles*.

2.0 Experimental and Computational Approach

Atmospheric aerosol particles were collected from an Amazon rainforest using a “Uniform Deposition Impactor” (cascade impactor). This method allowed particles of select sizes being deposited on a substrate for microscopy analysis. Scanning electron microscopy images were collected on the particles, where imaging was coupled with energy dispersive x-ray spectrometry to get the relevant chemical (elemental) information. The chemical information allowed for the analysis of mixing states, and the possible identification of different classes of fungal spores and bacteria. About 24,000 images with correlated elemental information were used as input for the machine learning step. Particle coordinates and dimensions from the microscopy images were correlated with elemental compositions; inorganic content from dust/minerals was characterized by the presence of the elements Al, Si, and Fe, while the organic component by the presence of C, N, O, and P. Elemental composition of each particle constituted the labels for prediction, evaluated both as continuous numerical and binary labels, for regression and classification tasks, respectively. To extract the particles of interest from the microscopy images, each particle was cropped and subsequently padded to uniform size (**Fig. 2a**). Analysis of particle size distribution led to the selection of input size 96 pixels in each dimension, as 98% of particles fell below this cutoff (**Fig. 2b**) and accommodating larger particles would needlessly increase computational complexity.

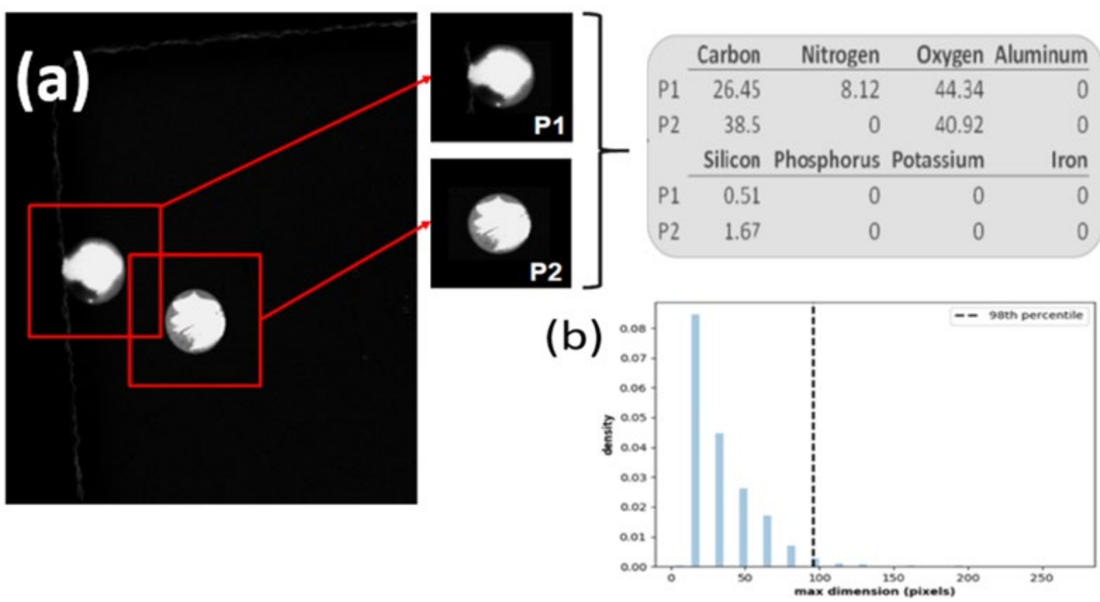

Figure 2: (a) To extract the particles of interest from the microscopy images, each particle was cropped and subsequently padded to uniform size. (b) 98% of particles fell below the 96 pixels cutoff.

3.0 Results

We have built a basic convolutional neural network (CNN, see **Fig. 3**) consisting of a series of alternating convolutional layers and max pooling layers, followed by a dropout layer, and two fully connected dense layers, the latter activated by a sigmoid function for binary prediction, or activated linearly for regression. Convolution involves a sliding kernel that passes over the input image in two dimensions, resulting in a new output "feature image". Multiple of these kernels are calculated simultaneously, each referred to as a "filter". In max pooling, the maximum value is taken for subregions of the input image, yielding a smaller output that contains said maximum value, enabling the network to generalize to differences in image orientation. Dropout involves randomly deactivating individual nodes of the network (here, 20%), but only during training. This discourages overfitting to the training data, and ultimately aids in the networks ability to generalize. Additionally, we implemented a modified version of Google's Inception v3 network, wherein the final layer is simply a fully-connected linearly- or sigmoid- activated dense layer, for regression and binary classification tasks, respectively, as opposed to a softmax output for many-class classification.

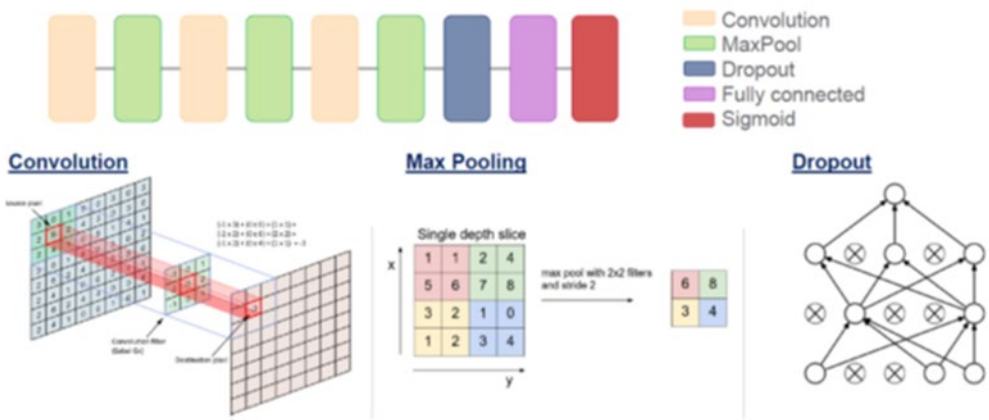

Figure 3: Schematic of the basic convolutional neural network (convnet).

Training involved withholding 33% of the data for validation, and training was performed for up to 1000 epochs (an epoch designates when the entire dataset is passed through the neural network once). The loss functions were mean squared error and binary crossentropy for regression and binary classification tasks, respectively. We selected adaptive moment estimation (Adam) as our optimizer, and used the AMSGrad variant, which involves an exponential moving average of the loss to perform weight updates. An early-stop criterion of 100 epochs was put in place to minimize overfitting effects – in essence, the network ceases training if validation loss does not improve for 100 epochs (**Fig. 4**). We additionally checkpointed the network to save the best-performing state in terms of validation loss. Finally, for regression, this resulted in mean absolute percent error of 22.26% for the basic CNN, and 20.17% for the Inception_v3-like network for organic atoms (C, N, O, and P), and a surprisingly high mean absolute percent error of 76.03% and 74.87% for the basic CNN and Inception_v3-like networks, respectively, for inorganic atoms (Al, Si, Fe). The significantly higher error for inorganic atoms is the focus of further investigation. For binary classification, the basic CNN achieved an accuracy of 84.29% across all atom types, and the Inception_v3-like network achieved an accuracy 85.51%. In all, the additional complexity introduced by Inception_v3 did not net significant improvement, indicating that either more data is required, or that performance was saturated with this image modality (that is, additional layers did not significantly impact results, such that the basic CNN may suffice).

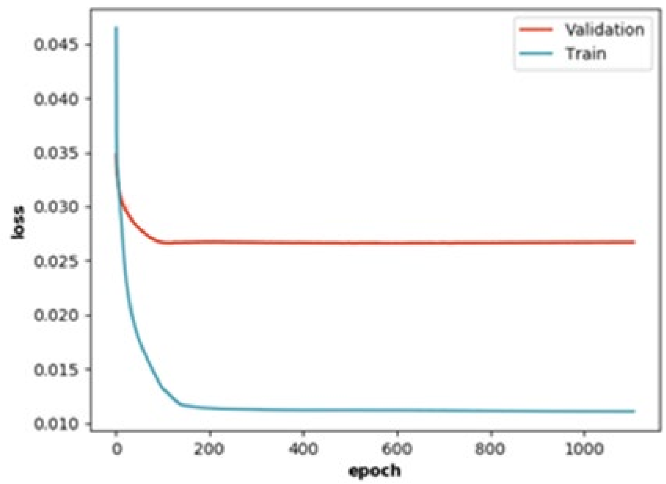

Figure 4: Training and validation loss vs the number of epochs. Final accuracy for binary classification for the basic CNN was of 84.29% across all atom types. The Inception_v3-like network achieved an accuracy of 85.51%.

The above platform will enable the efficient, streamlined analysis of thousands of particles by reducing analysis time, operator bias and error, and being more cost-effective. Our next steps in determining/predicting particle mixing states are: (1) to enhance deep learning with data input from x-ray imaging and x-ray absorption spectroscopy, (2) to experimentally verify our predictions, and (3) to introduce a multiscale aspect to this work. The latter involves the deep learning challenge of modeling the relationship between low-resolution image data and higher-resolution chemical information from spectroscopy. We also wish to extend the images to 3D data for component segmentation on soil aggregates/particles. As far as applications of this platform goes, we believe our network can be used for different data/ material systems as well as different instrumentation with small modifications to help users process large data sets. We have not responded to any funding solicitations, but we keep looking.

4.0 References

(1) Despres, V. R.; Huffman, J. A.; Burrows, S. M.; Hoose, C.; Safatov, A. S.; Buryak, G.; Frohlich-Nowoisky, J.; Elbert, W.; Andreae, M. O.; Poschl, U.; Jaenicke, R. Primary biological aerosol particles in the atmosphere: a review. Tellus B 2012, 64.

(2) Elbert, W.; Taylor, P. E.; Andreae, M. O.; Poschl, U. Contribution of fungi to primary biogenic aerosols in the atmosphere: wet and dry discharged spores, carbohydrates, and inorganic ions. Atmos Chem Phys 2007, 7, 4569-4588.

(3) Grote, M.; Valenta, R.; Reichelt, R. Abortive pollen germination: A mechanism of allergen release in birch, alder, and hazel revealed by immunogold electron microscopy. J Allergy Clin Immun 2003, 111, 1017-1023.

(4) Taylor, P. E.; Flagan, R. C.; Valenta, R.; Glovsky, M. M. Release of allergens as respirable aerosols: A link between grass pollen and asthma. J Allergy Clin Immun 2002, 109, 51-56.

(5) Huffman, J. A.; Sinha, B.; Garland, R. M.; Snee-Pollmann, A.; Gunthe, S. S.; Artaxo, P.; Martin, S. T.; Andreae, M. O.; Poschl, U. Size distributions and temporal variations of biological aerosol particles in the Amazon rainforest characterized by microscopy and real-time UV-APS fluorescence techniques during AMAZE-08. Atmos Chem Phys 2012, 12, 11997-12019.

(6) Poschl, U.; Martin, S. T.; Sinha, B.; Chen, Q.; Gunthe, S. S.; Huffman, J. A.; Borrmann, S.; Farmer, D. K.; Garland, R. M.; Helas, G.; Jimenez, J. L.; King, S. M.; Manzi, A.; Mikhailov, E.; Pauliquevis, T.; Petters, M. D.; Prenni, A. J.; Roldin, P.; Rose, D.; Schneider, J.; Su, H.; Zorn, S. R.; Artaxo, P.; Andreae, M. O. Rainforest Aerosols as Biogenic Nuclei of Clouds and Precipitation in the Amazon. Science 2010, 329, 1513-1516.

(7) O'Sullivan, D.; Murray, B. J.; Ross, J. F.; Whale, T. F.; Price, H. C.; Atkinson, J. D.; Umo, N. S.; Webb, M. E. The relevance of nanoscale biological fragments for ice nucleation in clouds. Sci Rep-Uk 2015, 5.

(8) Steiner, A. L.; Brooks, S. D.; Deng, C. H.; Thornton, D. C. O.; Pendleton, M. W.; Bryant, V. Pollen as atmospheric cloud condensation nuclei. Geophys Res Lett 2015, 42, 3596-3602.

(9) Guyon, P.; Graham, B.; Roberts, G. C.; Mayol-Bracero, O. L.; Maenhaut, W.; Artaxo, P.; Andreae, M. O. Sources of optically active aerosol particles over the Amazon forest. Atmos Environ 2004, 38, 1039-1051.

(10) Wikipedia https://en.wikipedia.org/wiki/Soil.

Appendix A – Products from the project

Publications:

(1) “Machine Learning Approaches for Analysis of Multiscale Imaging Data for Atmospheric and Soil Particles”, S China, S Colby, AK Battu, T Varga, Microscopy and Microanalysis, 25 (S2), 194-195 (2019);

Presentations:

(1) Varga T., S. China, S.M. Colby, and A. Battu, "Machine Learning Approaches for Analysis of Multiscale Imaging Data for Atmospheric and Soil Particles", PNNL TechFest, June 6, 2019;

(2) S. China, S. Colby, A.K. Battu, T. Varga, “Machine Learning Approaches for Analysis of Multiscale Imaging Data for Atmospheric and Soil Particles”, Microscopy and Microanalysis, August 4-8, Portland, OR;

(3) “Analysis of Internally Mixed Primary Biological Aerosol and Soil Particles using Machine Learning Approaches”, T. Varga, S. Colby, A.K. Battu, S. China, to be presented as poster and flash talk at EMSL Integration 2019, October 8-10, PNNL.

Capability developed:

Code “Deep convolutional neural network for particle characterization” uploaded to GitHub under https://github.com/pnnl/particle-net

Pacific Northwest
National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov