Applying machine learning (ML) and statistical models to low-cost aerosol sensors for anomaly detection

The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

Approved for public release; distribution is unlimited.

Chase Baker, John Helmsen , Ph.D., William Kaczynski, Ph.D., Sean Kinahan, Ph.D., Cody Rutherford, Ahmad Said, Nathan Spivy, Riley White, **Justin Taylor, Ph.D.**
Noblis, Reston, Virginia, USA

INTRODUCTION

The United States is investing in multiple programs to improve capability to monitor for airborne releases of chemical or biological agents using sensors and anomaly detection data analytics. For example, the Biological Detection for the 21st Century (BD21) program invested in sensors to collect particle size, concentration, and fluorescence data, which are analyzed using data fusion techniques and ML algorithms to distinguish biological threats from background particles (Bryan and Richardson, 2020). Despite significant investments, the Government Accountability Office (GAO) has stated, **“Biological aerosol sensors that monitor the air are to provide data on biological material in the environment, but common environmental material such as pollen, soil, and diesel exhaust can emit a signal in the same range as a biological threat agent, thereby increasing false alarm rates… false alarms produced by biological sensor technologies could be reduced by using an anomaly detection algorithm in addition to the sensor”** (GAO, 2021).

Recent developments in ML techniques offer the ability to move from traditional classification of detection events at individual sensors, to an approach fusing data from multiple sensors, weather stations, and an understanding of local patterns and sources that may vary diurnally, seasonally, or weekly. This research compared multiple approaches for biological aerosol anomaly detection using high density, low-cost, particle counters distributed throughout an urban environment.

METHODOLOGY

Our approach utilized data from PurpleAir sensor network (**Figure 1**) available at **www.purpleair.com**. These low-cost, networked aerosol spectrometers (**Figure 2**) provide data across six size bins, and we acquired two years of aerosol, and supporting sensor data from the Washington, D.C. metro region:

- 104 PurpleAir sensors measuring in 0.3, 0.5, 1, 2.5, 5, and 10 µm size bins, providing particle concentration per deciliter, temperature, and humidity every two minutes
- Meteorological data on wind speed and direction at hour intervals

We developed synthetic data by modeling aerosol transport of a 50 kg release of 1 µm anthrax spores once per month for all 24 months of data using the Hazard Prediction and Assessment Capability. Simulated data, converted to particles/deciliter, was overlaid with existing PurpleAir data.

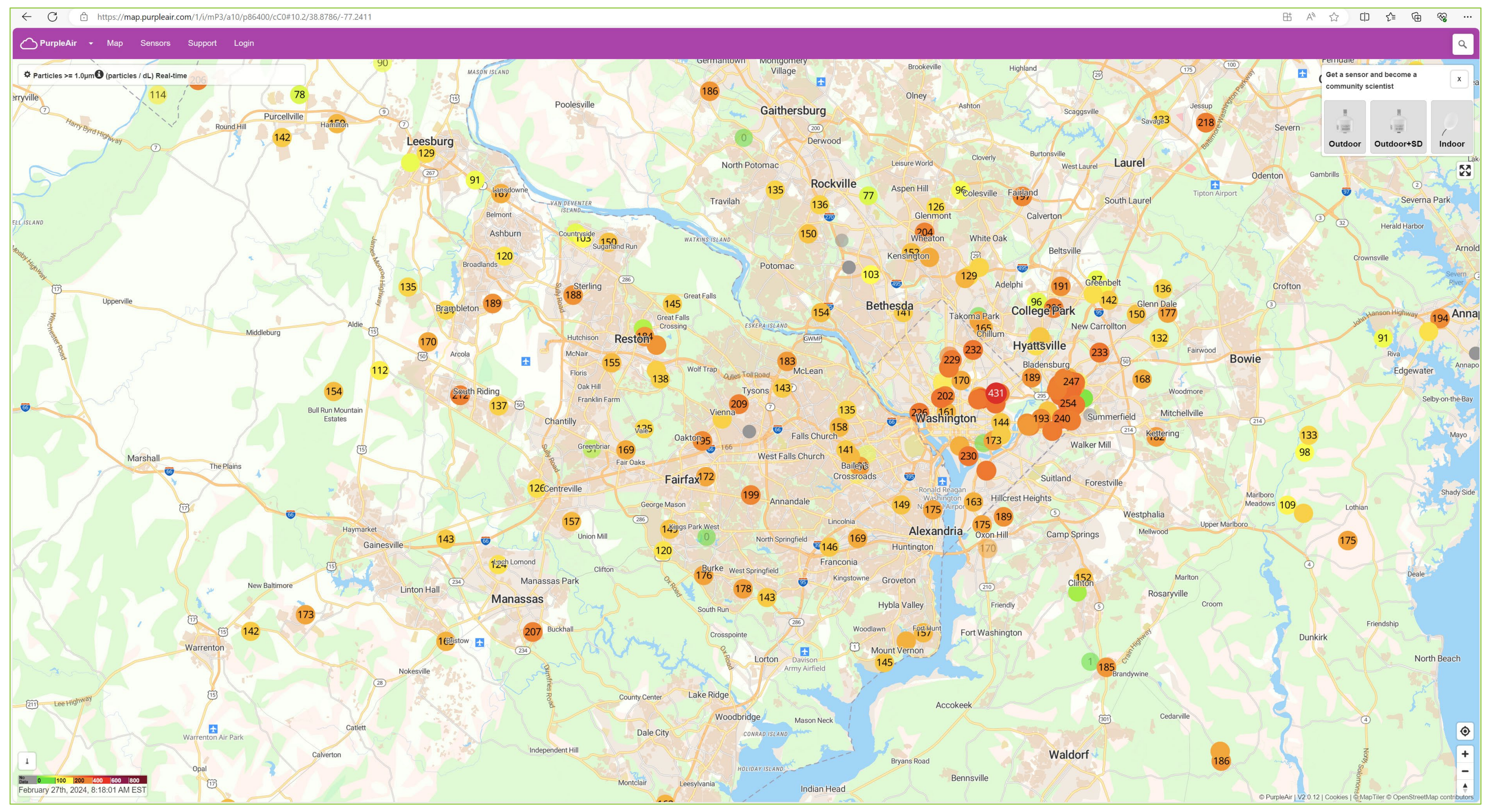

Figure 1. PurpleAir outdoor sensors measuring the concentration of 1 µm particles in the Washington, D.C., metro region.

Figure 2. PurpleAir Flex Air Quality Monitor.

ML APPROACHES

Autoencoders compress data into a lower-dimensional representation and reconstruct it back to its original form. Autoencoders can identify anomalies in time series because of reconstruction error and minimize the difference between the input and its reconstruction, which leads to learning the normal patterns. With new data inputs, data points that significantly deviate from learned patterns (i.e., have high reconstruction errors) are identified as anomalies.

Convolutional Neural Networks (CNNs), traditionally used for image processing, are also effective for time series data due to their ability to extract high-level features through hierarchical layers of convolutions. Each layer applies filters to the input data which capture temporal dependencies and feature hierarchies. Anomalies are detected by assessing deviations from the learned feature norms.

Gradient-Boosted Decision Trees work by sequentially building an ensemble of weak models, usually decision trees, where each subsequent model attempts to correct the errors made by the previous one.

Graph Neural Networks (GNNs) utilize the spatial relationships within graph-structured data, to capture complex patterns and dependencies. Anomalies are detected by examining deviations from the typical patterns learned across the graph. The inherent natural noise in particle size readings makes identifying an anomaly at a single sensor challenging, which is why this approach is favorable.

K-Means Clustering is a ML model that tags a dataset with a statistically significant set of cluster groups. Once a set of clusters have been assigned to a dataset, patterns emerge that can be used to provide deeper insights into the data under analysis.

Long Short-Term Memory (LSTM) utilizes gates that manage information flow over extended periods, enabling effective learning of long-term dependencies in time series data, improving upon traditional recurrent neural networks.

Transformer models leverage self-attention mechanisms to process sequences in parallel, rather than sequentially, enhancing efficiency and scalability. This architecture excels in understanding relationships within data, which is crucial for complex data.

STATISTICAL APPROACH

Benford’s Law predicts that in naturally occurring datasets, the leading digit "1" will appear about 30.1% of the time, "2" around 17.6%, and the frequency decreases logarithmically up to "9". To check for conformity, the observed frequency of each leading digit in sample data is compared to these expected proportions using statistical tests such as the chi-square goodness-of-fit test. If the underlying data's leading digit distribution conforms to Benford's Law and a newly obtained sample significantly deviates from Benford's expected distribution, it suggests that the data may not be naturally occurring or that an anomaly has occurred.

RESULTS

The various ML approaches and statistical approach were developed/trained and tested on the data set. Two of the ML approaches demonstrated initial success (**Table 1**), while the remaining four ML approaches and statistical approach were unsuccessful despite a large release (50 kg) used in the synthetic data.

- The successful ML approaches were Gradient-Boosted Decision Trees, and Transformer Models
- The unsuccessful ML approaches were Autoencoders, Graph Neural Networks, K-means Clustering, and Long Short-Term Memory
- The unsuccessful statistical approach was Benford’s Law

Table 1. F-Score, Precision, and Recall of ML approaches. F1 is the combined form of precision and recall. Precision is number of true predicted positives / number of total predicted positives. Recall is number of true predicted positives / number of all ground truth positives.

Type	Model	F-score	Precision	Recall
Decision Tree	**XGBoost**	**0.591**	**0.565**	**0.619**
Transformer	**MEMTO**	**0.2842**	**0.2367**	**0.3555**
Transformer	Donut	0.1049	0.0924	0.1214
LSTM	EncDecAD	0.0426	0.0223	0.4786
CNN	SRCNN	0.0366	0.0190	0.4786
Graph	FuSAGNet	0.0279	0.0143	0.6412
LSTM	LSTMAD	0.0245	0.0124	0.9
Graph	GDN	0.0172	0.0088	0.3358
AutoEncoder	AutoEncoder	0.0125	0.5745	0.0244

CONCLUSIONS

Overall, smarter data fusion is expected to play a larger role in biological incident detection and response, detect a biological release, map the extent of contamination, predict potential consequences, and enable a response. **Due to the high variability in background aerosol concentration, anomaly detection can’t serve as a standalone detection modality but can support detection if integrated with the appropriate concept of operations.** As ML algorithms improve, low-cost, high density, distributed aerosol spectrometers are likely able to assist in anomaly detection, but data variability will continue to be a limited factor.

REFERENCES

Bryan, B.N. and Richardson, D.E. (2020) *DHS Biosurveillance Systems.* Fiscal Year 2020 Report to Congress. https://www.dhs.gov/sites/default/files/publications/st_cwmd_-_dhs_biosurveillance_systems.pdf

Government Accountability Office. (2021) *DHS Exploring New Methods to Replace BioWatch and Could Benefit from Additional Guidance*. Report to Congressional Requesters. https://www.gao.gov/assets/720/714434.pdf

Learn more at: noblis.org/counterWMD