INNOVATING CROSS-DOMAIN SOLUTIONS TO DETECT EMERGING BIOLOGICAL THREATS

Applying Machine Learning And Statistical Models To Low-cost Aerosol Sensors For Anomaly Detection

Justin Taylor Noblis **Brice Ballesteros** Noblis **John Helmsen** Noblis **Sean Kinahan** Noblis **Oscar Olmedo** Noblis **Cody Rutherford** Noblis **Ahmad Said** Noblis **Nathan Spivy** Noblis **Riley White** Noblis

Currently, the U.S. is investing in homeland biodetection architectures that perform monitoring for airborne releases of biological agents using anomaly detection sensors and data analytics. The sensors collect particle size, concentration, and fluorescence data, which are analyzed using data fusion techniques and machine learning algorithms to identify biological threats from background particles. Despite significant investments, the Government Accountability Office 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.”

Low-cost aerosol sensors (that measure particle size and concentration) deployed by individuals in populated areas have recently gained popularity. PurpleAir (https://map.purpleair.com/) has a real-time air quality map with sensor data on particle size distribution and concentration, as well as an application programming interface that facilitates data collection. We hypothesize that the greater concentration of PurpleAir sensors may enable development of advanced anomaly detection algorithms that overcome the current challenges with background aerosols. To test our hypothesis, we have identified and developed various statistical and machine learning approaches to identify anomalies within PurpleAir datasets.
We have ingested two years of PurpleAir data from outdoor sensors across the Washington, D.C., metro region. Data consist of particle concentrations across 0.3 µM, 0.5 µM, 1.0 µM, 2.5 µM, 5 µM, and 10 µM size ranges, as well as temperature and relative humidity. Additionally, we have obtained the meteorological data on wind speed and direction for the relevant aerosol sensor measurements. We've engineered an array of statistical and machine learning methodologies for anomaly detection within temporally aggregated data. Our approach uncovers irregularities and inconsistencies across geographic data by analyzing data patterns, time trends, relational structures, and sequential deviations. We developed synthetic data with various meteorological conditions to simulate anomalous events by replicating the particle size distribution and concentration within a biological plume and overlaying the data on existing PurpleAir datasets with matching meteorological conditions.

We will present the different algorithm approaches that we have developed with the benefits and disadvantages of each approach in identifying the synthetic anomalous events that represent a biological attack. This includes the sensitivity, specificity, timeliness, and computational requirements for each approach. We have also identified requirements for sensor density and placement for optimal detection as additional sensors increase sensitivity and specificity but incur an additional cost and require more resources. This study informs anomaly detection development to protect the Joint Force from aerosol threats.