**Project Title:** Using computer vision to identify onsite wastewater systems

**Background**

The field of Environmental Science has a long-standing history of using remote monitoring approaches to assess a wide variety of conditions that include weather, geologic, and man-made features. (lit review needed) Computational capacity has increased significantly in recent years, allowing the development of a number of novel approaches to analyzing and interpreting digital images, a process some now labeled as “computer vision.” Machine learning approaches, particularly neural networks, can use the relationships between groups of values to develop models that identify features. The most common current approaches in the area of computer vision use supervised learning. This approach “trains” a neural network by processing (often many times) a set of images with features that have been labeled by human subject matter experts until it is able to identify these features with some degree of accuracy. A separate set of images that have also been labeled, but the model has not yet analyzed, is then evaluated by the model to assess the accuracy of the model in identifying these features. Iterative approaches to train and test the model can be used to improve the accuracy and evaluate its performance across a variety of images.

While computer vision-related work has a long history of use in environmental science (often labeled as “remote sensing”) there has been very little use of computer vision in environmental public health to-date. Staff at the Centers for Disease Control and Prevention (CDC) and the Agency for Toxic Substances and Disease Registry (ATSDR) have piloted the use of computer vision approaches to identify water cooling towers in metropolitan areas. This project was intended to be used by public health staff during investigations of legionnaires disease or Pontiac Fever so they could more readily identify water cooling towers, a common source of the *Legionella* bacteria that cause these diseases. ([reference CDC](https://www.cdc.gov/legionella/index.html)) The project achieved a 90% accuracy and continues to be evaluated and improved.

Traditional statistical approaches using multi-variable models to predict the presence or absence of a system from a set of variables as well as the system’s potential functionality (i.e., failed versus functioning) may also assist jurisdictions with this work. Examples of these approaches include xxx, xxx, xxx.

Researchers from the CDC, ATSDR, and the Environmental Protection Agency (EPA) engaged federal, state, and local environmental protection and health agencies to assess the past or present use of traditional statistical and computer vision approaches to identify onsite wastewater systems. The outreach included an assessment of any realized or potential benefit to improve environmental public health practice. These discussions led to the decision that there was utility to both the field and environmental public health practice in exploring this topic in more depth. The group of experts determined that piloting this work with multi-family onsite wastewater systems (aka “package systems”) would provide the greatest chance of success and a further proof of concept for this approach.

The intent of this project is to explore the utility in applying traditional multivariable models and supervised machine learning approaches to identify onsite wastewater systems (aka septic systems). The primary focus will start with accurate identification of a septic system with a secondary evaluation of whether it would be possible to distinguish currently-functioning systems from failed ones. Initial efforts will start with identification of multi-family wastewater systems (aka “package systems”) with potential later work on individual household systems. If successful, this approach will greatly relieve the resource requirements for CDC and jurisdictional programs to accurately identify onsite wastewater systems. This work will also advance this field’s use of this potentially powerful approach for use in other public health issues.

**Methods**

This study underwent the CDC Human Subjects Review process and was considered exempt from review as it used publicly-available information without any personal identifiers.

Placeholders for methods topics

* Data was collected from (LA St. Tammany parish, others)
* Time period of data used
* Traditional statistical methods used
  + Variables
* CV approaches used (CNNs)
  + How many SMEs need to identify a system (maybe 2 independent with a 3rd person in a meeting to review and adjudicate any discrepancies?)
  + How to identify known failed systems to label.
  + Time periods needed
  + Vegetation indices to evaluate. Potential citation

Next steps, updated 5/9/24

* Focus on St. Tammany to start
  + Potential other partners:
    - NC DOH
    - RI DOH
    - GA Dept. of EH
    - [Massachusetts Alternative Septic System Test Center](https://www.masstc.org/)

**Anticipated results**

Table 1.

* Total # of known systems vs. # with sufficient data to evaluate
  + Sub columns for traditional statistical model and CNN models
* Univariable stats

Table 2. Bivariable

Table 3. Multivariable (traditional) stats results

Table 4. CNN results

* Sensitivity / specificity for any system
* Sensitivity / specificity for failed system

Figure 1. Example images from CNN work

* Example of an identified system
* Example of something that looks like a package system but is not

Discussion

This project evaluated the use of traditional and novel approaches to identifying