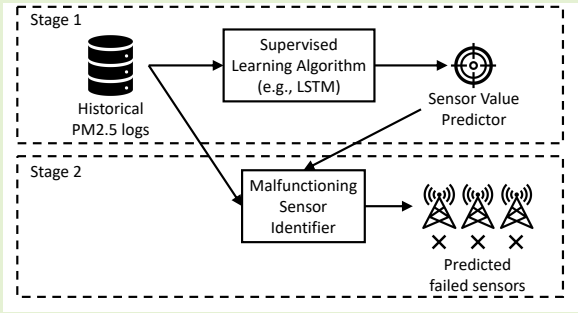
Learning to Identify Malfunctioning Sensors in a Large-Scale Sensor Network

Tzu-Heng Lin, Xin-Ru Zhang, Chia-Pan Chen, Jia-Huei Chen, and Hung-Hsuan Chen

***Abstract*— This paper proposes a two-stage methodology to discover malfunctioning sensors in an air quality sensor network. The two-stage methodology consists of a supervised learner to predict the future PM2.5 values of each sensor and a detector that leverages the result of the previous stage to detect the malfunctioning sensors. Consequently, even if each sensor's health status (i.e., normal or malfunctioning) is unavailable, we can still apply powerful supervised learners to this task. We conduct experiments on a nationwide air quality sensor network that includes 10,000+ sensors and utilize periodic maintenance records on some of these sensors as the ground truth of their health status. Experimental results show that this two-stage methodology can effectively discover problematic sensors. As maintaining a large-scale sensor network is laborious, the methodology can dramatically reduce the human resource required for regular inspection.**



***Index Terms*— auto inspection, IoT, AIoT, PM2.5, sensor network, anomaly detection**

I. Introduction

A Sensor network contains a large number of sensors that collaboratively monitor and collect information from a target environment. In an ideal situation, each sensor detects only local information and sends information to neighbors, local servers, or a centralized server. The small pieces of collected information are integrated to obtain a large picture of the target. This concept has been applied to various applications, such as environmental monitoring (e.g., detecting regional temperature and humidity) [1], industrial monitoring (e.g., detecting machine temperature and vibration frequency) [2], healthcare monitoring (e.g., detecting pulse rate and oxygen saturation) [3], and many more.

When the number of sensors is large enough, we sometimes assume that the failure of a small portion of sensors can be ignored or corrected since the measurements from the remaining healthy sensors may still be sufficient to assemble the global picture [4]. However, this assumption is valid only when a system provides certain degrees of fault tolerance [5]. In practice, malfunctioning sensors may still transmit incorrect monitored values to a centralized server or to neighboring nodes. When these measurements contain more noise than information, integrating these measurements into a system may bring more harm than benefit. Consequently, it is essential to identify malfunctioning sensors and fix problems rapidly.

This paper presents a study on determining malfunctioning sensors from a large-scale air quality sensor network. We use the terms "malfunctioning sensors", "failed sensors", and "abnormal sensors" interchangeably in this paper. The large-scale sensor network studied in this paper is deployed by the Environmental Protection Administration (EPA) of Taiwan. This network contains 211 large stations to monitor specially selected locations and $10,000+$ middle-sized air pollution sensors in 282 townships and 111 major industrial and science parks in Taiwan. Most sensors are located on the west coast of Taiwan because the west coast is the most populated (90% of 23 million people[1]) and contains major industrial areas that produce most air pollution.

Determining the malfunctioning sensors from this sensor network is an exciting and vital task for several reasons. First, this is a live sensor network containing approximately $10,000$ working sensors in the area of 36,000+ km^2 for several years. A method that works in this scope may prove its practicability in a large-scale sensor network. Second, some of these sensors may report inaccurate values, but we do not know which ones a priori. Many sensor networks may also face a similar situation, especially those with a large number of sensors distributed in

This work was supported in part by the Ministry of Science and Technology in Taiwan under grant number MOST 110-2222-E-008-005-MY3 and MOST 110-2634-F-008-008.

T.-H. Lin and X.-R. Zhang were with the Department of Computer Science and Information Engineering (CSIE), National Central University (NCU), Taiwan. H.-H. Chen is with the Department of CSIE at NCU. E-mails: (neilbeeber@gmail.com, appleisgoodgood-eat@gmail.com, hhchen@g.ncu.edu.tw).

C.-P. Chen and J.-H. Chen are with the Industrial Technology Research Institute, Taiwan. E-mails: (peter_chen@itri.org.tw, 60523001l@gapps.ntnu.edu.tw).

Corresponding author: Hung-Hsuan Chen.

[1] https://en.wikipedia.org/wiki/Demographics_of_Taiwan

a wide area. Third, we obtained several ground truth labels (i.e., a sensor is normal or abnormal) from on-field inspections by comparing the monitored values of a sensor with those reported by a lab-calibrated device with high accuracy. Consequently, we can compare the predicted results with the ground truth status of the inspected sensors. In contrast, most previous studies examining malfunctioning sensors did not have ground truth labels and usually made simplified assumptions (e.g., labeling outdoor sensors as normal and indoor sensors as abnormal) to infer the status of each sensor.

The rest of the paper is organized as follows. Section II outlines previous studies on air quality sensors, autonomous measurement calibration, and failed node detection. Section III presents our methodology to identify the suspected malfunctioning sensors. Section IV introduces the experimental dataset and compares our proposed model with several baseline methods. Finally, we discuss the discovery, limitations of our method, and ongoing and future work in Section V.

II. Related Work

This section reviews previous works on air quality sensors, applications related to air quality sensors, and studies on detecting malfunctioning sensors.

A. Air Quality Sensors and Applications

Low-cost air quality sensors have become a fundamental infrastructure to monitor air quality on a large scale with fine granularity [6]–[9]. However, this infrastructure also presents many new challenges, such as the best practices for sensor deployment, sensor calibration, data management, and data integration. Low-cost sensors sometimes report very different measurements to devices that employ more advanced monitoring approaches, e.g., federal reference methods (FRMs) and federal equivalent methods (FEMs) [6]. Consequently, low-cost sensors may require periodic maintenance and calibration so that data quality can be assured. The strategy of deploying many low-cost sensors brings an apparent trade-off: while we can install many sensors at an affordable price, the maintenance costs can be excessive.

Many studies have proposed utilizing monitored PM2.5 values to predict future PM2.5 values or future trends. Most of these studies suggest utilizing supervised learning models to find the relationship between the features (mainly including the previously monitored values) and the targets (future observed values) [10]–[13]. Among them, the linear regression model is simple and highly interpretable because each feature is assumed to have a linear relationship with the target variable [14]. However, this model cannot capture the nonlinear relationship between the target and the features. The decision tree and the random forest models also have good interpretability, but integrating the spatial or temporal relationship is not straightforward [14]. Some works apply spatio-temporal interpolation, e.g., kriging, to predict the value of a sensor based on spatial and temporal information (e.g., [15], [16]). However, kriging requires the covariance matrix, which is computationally expensive when the input has large dimensions. To efficiently integrate spatial, temporal, and perhaps other types of features among sensors, advanced machine learning models, such as the support vector machine and various deep learning models, are applied to capture the nonlinear relationships among the features and the target variable [17]–[19]. While complex supervised learning models may be flexible in integrating various information, these models assume that previous monitored values are correct, which might be an overly naïve assumption that could impede the accuracy of future PM2.5 prediction. Our model in Stage 1 (details in Section III-E) also suffers from the same issue. However, our objective – identifying the malfunctioning sensors – differs from the above-mentioned works (predicting future PM2.5 values). As a result, the imperfection of the previously monitored PM2.5 values may cause imperfect future PM2.5 prediction, which on the contrary, helps our model in Stage 2 to identify the problematic sensors (details in Section III-F).

Many papers have leveraged the output of air quality sensors as the input of downstream tasks, e.g., identifying and localizing pollution sources [20], querying air quality through a smart interface (e.g., via a chatbot [21]), detecting unusually high concentrations of air pollutants [22], and many more. But, again, the foundation of these downstream applications is the correctness of the monitored values. Unfortunately, this is sometimes an unrealistic assumption.

B. Malfunctioning Sensor Detection

To ensure the current status of low-cost air quality sensors, one evident approach is conducting periodic inspections. Unfortunately, this is a laborious process, especially when many sensors are distributed in a large area.

Many previous works have studied methodologies to discover problematic sensors; some have even suggested calibrating the monitored values automatically or semiautomatically [23]–[30]. The most straightforward strategy to find an abnormal sensor is using rules and simple statistics to select the outliers as the malfunctioning sensors [31]. Apparently, human-defined rules are limited by rule-makers' knowledge and expertise. As a result, many studies suggest utilizing data-driven approaches. When the ground truth label of normal or abnormal is unavailable as the target for training, we usually need to rely on various unsupervised approaches such as clustering [23], [32], [33], principal component analysis (PCA), or kernel principal component analysis (KPCA) [26], [34], [35]. On the other hand, if the ground truth label is accessible as the training data, we can utilize various supervised learning algorithms to train classifiers. Among the supervised learning models, the linear model is simple, fast, and easy to interpret [10], [11]. However, a linear model is unlikely to discover the high-dimensional interactions among the heterogeneous features and the target variables. More complicated supervised learning models, such as support vector machines [11], random forests [11], artificial neural networks or deep learning [10], [12], or a combination of multiple supervised models [36] have also been employed on this problem to discover more complicated patterns. Surveys of the challenges and anomaly detection in the IoT and sensor network environments were given in [25], [37], [38]. Some

studies further investigated automatic calibration when sensors become less accurate. For example, CalibrationTalk calibrates inaccurate sensors caused by aging [30]. Kriging was used previously to calibrate sensors. However, likely due to the high computational cost, it was applied to sensor networks with only dozens of sensor stations [13]. One can also apply supervised learners to predict future monitoring values and use the predicted values for calibration [24], [27], [28], [39].

Although all these models aim to model the relationship between the sensor status (i.e., normal or malfunctioning) and various clues (e.g., the monitored PM 2.5 values from the neighboring sensors, the previous monitored values from the target sensor, etc.), they mostly exhibit some of the following problems. First, many relevant studies do not have ground truth on the sensor status (normal or malfunctioning) during evaluation. Consequently, the reported performance is questionable. Second, even if some papers rely on simple heuristics to infer sensor status, these inferences may not necessarily be correct, and the number of the inferred training instances is limited, so the corresponding models probably overfit the training data and are hard to apply in a real environment. [6]. For similar reasons, we choose not to use these datasets in our experiments.

Our study involves a rigorous data collecting policy and a unique two-stage learning strategy to overcome the above two issues. First, we obtain the ground truth of the sensor status by onsite inspections (details are introduced in Section III-A). Second, our experimental dataset involves 12 months of monitoring logs from 144 sensors, and the training involves only monitoring values but not the sensor status. Consequently, we have a large number of training instances.

III. Design

This section presents the current status of Taiwan's air quality sensor network, the problem definition, and the methodology of identifying malfunctioning air quality sensors.

A. Air Quality Sensor Network in Taiwan

Environmental protection is an essential issue in the modern world. Among these issues, air pollution and PM2.5 are undoubtedly important aspects. PM2.5 refers to atmospheric particulate matter (PM) with a particle size less than 2.5 micrometers (μm). Because PM2.5 is tiny, mucous membranes or cilia cannot block it from entering the human body. Additionally, the composition of PM2.5 is highly complex and may be attached to substances that could harm humans. It has been shown that exposure to PM2.5 for a long time increases the chances of getting cancer. The World Health Organization (WHO) has reported that an average annual concentration of PM2.5 above $10\mu g/m^3$ can affect human health [40].

To effectively monitor the values of PM2.5 in different areas, the Environmental Protection Agency (EPA) of Taiwan has established 211 highly accurate national-level monitoring stations since 1998 to assess the air quality over a wide range, and the EPA regularly calibrates the sensors of these stations to ensure the correctness of observations. These stations can measure highly accurate PM2.5 values, but the number of stations is low due to the high cost. To monitor regional air pollution hotspots with a finer granularity at an affordable cost in real-time, the EPA has set up $10,000+$ less expensive regional optical PM2.5 sensors nationwide to form an "air quality sensor network" in an area of $36,000+$ km^2 (the size of Taiwan). These low-cost sensors have been gradually established since 2017. In addition to the PM2.5 values, each sensor also detects the temperature and the relative humidity. This study targets these less expensive sensors. The historical monitoring dataset has been released and can be downloaded from `https://ci.taiwan.gov.tw/dsp/dataset_air.aspx`.

Although our dataset contains not only the PM2.5 values but also the humidity and temperature that may be helpful to predict future PM2.5 values, these extra features may not always be available in other sensor networks. We demonstrate that our model is effective even using only the previous PM2.5 values as the input features (details in Section III-E). Therefore, other air quality sensor networks can apply our two-stage model even if these sensor networks only collect PM2.5 values. That being said, we still tested the effectiveness of these features in the experiments.

B. Challenges

An optical sensor determines the PM2.5 values by drawing the particles in the air into the sensing area and measuring the numbers of particles of different sizes via light scattering. Although optical sensors are carefully examined in the onset, they may gradually become less accurate for various reasons. First, due to the differences in particle diameters, shapes, surface roughness, and other physical properties, light may have different degrees of reflection and scattering, which may influence detection accuracy. Second, the fan motor to draw the air can only sample a small portion of air at a time, so the measured PM2.5 values may vary dramatically within a short period, especially when the air pollution is distributed unevenly. Third, certain chemical materials (e.g., sulfate and nitrate) in the particles may absorb water, which may cause the particles to deform. Consequently, the relative humidity affects the monitored PM2.5 values. Fourth, environmental factors may sometimes affect the measurement. For example, sensors may sometimes be hidden by trees, signboards, or other obstacles. Finally, and probably the most apparent reason, the sensors can become less accurate because of aging.

Currently, sensors are maintained by periodic inspections with a strict set of rules. Every season, the maintenance team samples sensors to conduct onsite inspections by comparing the sensors with a reference machine for at least 12 hours. The reported metrics include the relative error and coefficient of determination. The reference machine is calibrated by a national-level monitoring station before the inspection and verified again after the inspection. Since the number of sensors is large and the inspection process is laborious, maintaining these sensors comes at a significant cost. On the other hand, by reducing the maintenance frequency, the sensors' monitored values could become unreliable, further affecting the downstream applications that depend on the air quality data released

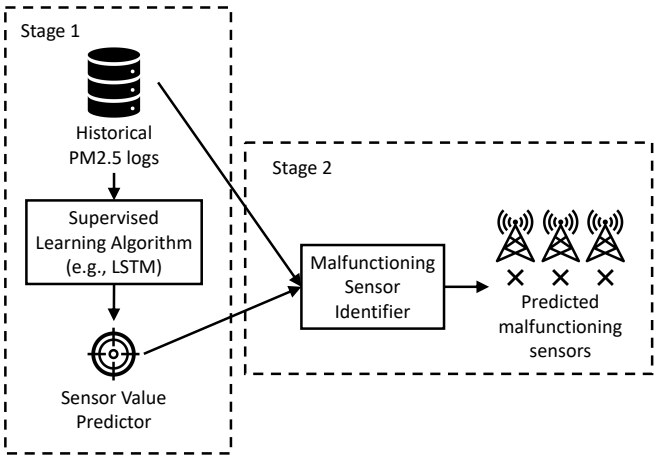


Fig. 1. An overview of the two-stage model detecting model

from the government. According to a maintenance record in 2018 (as we show later), the fault rate of a randomly sampled regional sensor is approximately 19.4%, suggesting that a good portion of the monitored PM2.5 values might be incorrect. Unfortunately, we do not know which ones are wrong.

C. Objective

We aim to use a data-driven mechanism to identify the suspicious malfunctioning air quality sensors to reduce the inspection and maintenance costs and increase the credibility of the monitoring values. At first glance, the problem may look like a standard binary classification task: given the features $\boldsymbol{x}_i$ of a sensor i and perhaps its neighboring sensors, predict whether this sensor is normal or malfunctioning. In practice, however, the labels of normal/malfunctioning are usually unavailable, i.e., we are generally unsure of the condition of a sensor. Consequently, it is challenging to apply supervised classifiers directly. Furthermore, even if some inspection results are available, the number of positive instances (i.e., the malfunctioning sensors) is usually small. For example, based on one of our inspection records involving 144 sensors in 2018, 26 sensors were malfunctioning. As a result, it is difficult for a classifier to identify the general pattern of a malfunctioning sensor based on these limited positive instances.

We design a workflow that leverages supervised learning models but requires no label of normal/malfunctioning as the target during training. Our proposed method can find the problematic regional sensors accurately. As a result, we can effectively reduce the maintenance costs by checking the sensor that our model identifies as problematic.

D. Proposed Method Overview

Instead of predicting the malfunctioning sensors directly, we propose a two-stage workflow to find the suspicious malfunctioning sensors. Figure 1 gives an overview of the two-stage model.

In the first stage, we use supervised learning algorithms to build a sensor value prediction module to predict the monitored PM2.5 value of each sensor in the near future. Since there are approximately 10 thousand sensors continuously monitoring the PM2.5 values, we have many training instances for this new task (although, in practice, some of the monitored PM2.5 values might be incorrect, so the target values and some feature values could be noisy). Once we obtain a good model to predict the measured PM2.5 values of the near future of a target sensor, we proceed to the second stage – identifying the sensors that have large residuals between the predicted PM2.5 values $\hat{y}_i$s and the monitored PM2.5 values (y_is). These should be sensors that are likely to be problematic. As such, we do not need to label each sensor as normal or malfunctioning during the training process, but we can still predict the malfunctioning sensors based on the supervised learning approaches.

The following two sections detail these two stages.

E. Stage 1: Sensor Value Prediction

To predict $\hat{y}_i^{(t)}$ the monitored value of sensor S_i at timestamp t, we utilize the spatial and temporal features.

Figure 2 shows how to generate the features and the targets based on the monitored PM2.5 values from the sensor station S_i. To predict $y_i^{(t)}$ the monitored value of S_i at time t, we generate features from two sources. The first source is from the target sensor: we use $y_i^{(t-1)}, y_i^{(t-2)}, \ldots, y_i^{(t-30)}$, the monitored PM2.5 values of the last 30 minutes from sensor S_i, as part of the features. The second source is from $n_{i,1}, \ldots, n_{i,5}$, the five closest sensors to sensor S_i: we include $y_{i,j}^{(t-1)}, \ldots, y_{i,j}^{(t-30)}$ $(j = 1, \ldots, 5)$, the monitored values of neighboring sensors $n_{i,j}$s, in the last 30 minutes as the features. Consequently, we denote a training instance in the first stage by $\left(\boldsymbol{x}_i^{(t)}, y_i^{(t)}\right)$, where $\boldsymbol{x}_i^{(t)}$ is denoted by Equation 1

$$\boldsymbol{x}_i^{(t)} = \left[y_i^{(t-1)}, \ldots, y_i^{(t-30)}, \boldsymbol{y}_{i,1}^{(t-1):(t-30)}, \ldots, \boldsymbol{y}_{i,5}^{(t-1):(t-30)}\right], \tag{1}$$

where $\boldsymbol{y}_{i,j}^{(t-1):(t-30)} = \left[\boldsymbol{y}_{i,j}^{(t-1)}, \ldots, \boldsymbol{y}_{i,j}^{(t-30)}\right]$ $(j = 1, \ldots, 5)$.

Among these features, $y_i^{(t-1)}, \ldots, y_i^{(t-30)}$ can be regarded as the temporal features (as they are the historically monitored values of the target sensor), $y_{i,1}^{(t)}, \ldots, y_{i,5}^{(t)}$ can be considered as the spatial features (as they report the neighboring monitored values at time t), and $y_{i,j}^{(t-u)}$ $(j = 1, \ldots, 5, u = 1, \ldots, 30)$ can be regarded as both the spatial and temporal features (because they represent the previously monitored values in the adjacent sensors). We rotate the value of t from 31 to the timestamp before the inspection day. Consequently, we can generate a large number of training instances.

We illustrate an example to further illuminate the process of generating the training data. In the 144 inspected sensors, the earliest inspection date is May 29, 2018, so we can create the training data based on the monitored value from day 1 to May 28, 2018 (the day before the inspection). If day 1 is Jan. 1, 2018, then we have 148 days to generate the training data, which corresponds to approximately $1440 \times 148 = 213,120$ training instances (as one day has 1440 minutes).

We experiment with various supervised learning models to predict the PM2.5 values in the near future. The predictors include traditional models (ridge regression with regularization

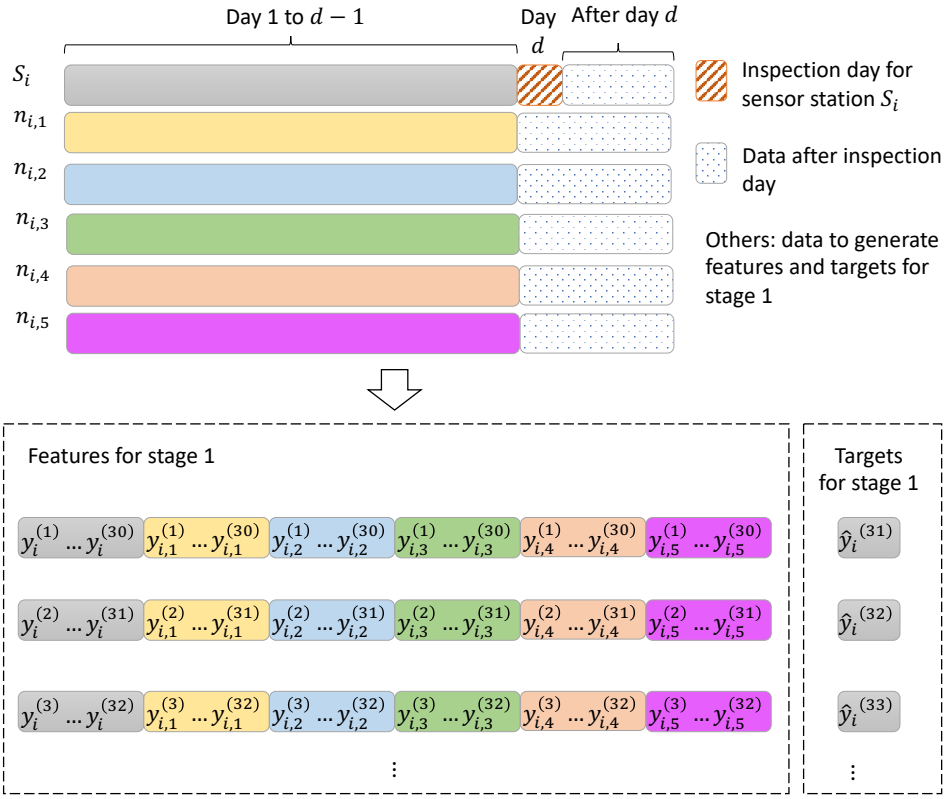


Fig. 2. Generating the labels (the monitored PM2.5 value of the next minute on a sensor $\boldsymbol{S_i}$) and features (the monitored PM2.5 values of $\boldsymbol{S_i}$ in the previous 30 minutes and the monitored PM2.5 values of the neighbors of $\boldsymbol{S_i}$ at the current moment and in the last 30 minutes).

weight 1.0, Lasso with regularization weight 1.0, and random forest with 25 trees) and deep-learning-based models (fully connected neural networks with 3 hidden layers whose sizes are 300, 100, and 20 and dropout rate 20%, and long short-term memory with 3 hidden layers, each with 128 neurons). We omitted kriging because it requires estimating a covariance matrix whose size is the square of the number of input samples. In our case, the matrix size is more than $200,000^2$. If each entry needs 4 bytes, storing the covariance matrix results in more than 150 GB, beyond the memory capacity of most modern desktop computers or even workstations.

One obvious drawback of such a method is that we assume the monitored PM2.5 values are correct. Since a good portion of the sensors could be problematic,[2] this assumption is overly optimistic. In particular, both the feature values (e.g., the PM2.5 values of the neighboring sensors) and the target values (i.e., the PM2.5 value we will predict for the centroid sensor) could be incorrect. However, we still apply these regression models directly for two reasons. First, as we have many training samples, the model is still likely to discover the relationship between the features and targets, even though some of the feature values or target values may be incorrect. Second, the final goal of this project is not to predict a perfect PM2.5 value for the near future but to discover the potentially malfunctioning sensors to reduce the manual inspection cost. If some features or targets are inaccurate, we may end up getting inaccurate predictions of the PM2.5 values for the corresponding sensors. Consequently, these sensors are likely to be regarded as malfunctioning sensors in the second stage.

[2]A regular random inspection in 2018 showed that 19.4% of the sensors are malfunctioning

F. Stage 2: Malfunctioning Sensor Detection (MSD)

We use the models mentioned in the above section to predict the monitored values of a sensor S_i at each minute on day d_i, the day we inspected sensor S_i. We call this stage MSD, which stands for Malfunctioning Sensor Detection.

Let the monitored PM2.5 values for S_i on day d_i be $\boldsymbol{y} = \left[y_i^{(1)}, \ldots, y_i^{(1440)}\right]$; we compare the predicted values $\hat{\boldsymbol{y}} = \left[\hat{y}_i^{(1)}, \ldots, \hat{y}_i^{(1440)}\right]$ with $\boldsymbol{y}$ based on the coefficient of determination (i.e., R^2 score). Each $\hat{y}_i^{(t)}$ is predicted based on the models with feature set $\boldsymbol{x}_i^{(t)}$ $(t = 1, \ldots, 1440)$ that were introduced in Section III-E.

The R^2 score can be defined by Equation 2.

$$R^2(\hat{\boldsymbol{y}}, \boldsymbol{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{2}$$

where $\bar{y}$ is the average of the targets in the training data and $n = 1440$.

We use the R^2 score instead of other popular metrics (e.g., root-mean-squared error or absolute error) because the R^2 score is a normalized score with the largest possible value of 1.

If using nonnormalized metrics, it could be difficult to interpret the goodness of a prediction based on a single number.

After obtaining the R^2 score for each sensor, we rank the sensors based on their R^2 score from the smallest to the largest and select the top k sensors with the smallest R^2 score as the malfunctioning sensors.

IV. Experiment

This section presents the statistics of the experimental dataset, the data preprocessing steps before conducting the experiments, and the experimental results for the Stage 1 (monitored PM2.5 value prediction) and Stage 2 (malfunctioning sensor prediction) tasks.

A. Experimental Dataset

The experimental dataset consists of two parts. The first part is the log of the monitored PM2.5 values released by the EDA. The second part is an inspection record on selected regional sensors in Taichung City, the largest city in central Taiwan.

The first part of the dataset, the monitored PM2.5 values from $10,000+$ sensors, was released by the EPA of Taiwan. This dataset is released daily starting in May 2017 and is still actively updated when writing this paper. This dataset contains the station IDs, the locations (i.e., the latitudes and longitudes) of the stations, historical PM2.5 values of each sensor (most sensors record the PM2.5 values per minute), and historical temperature and humidity. The EDA also provides APIs for developers to query the latest PM2.5 values of the sensors with a 3-minute updating frequency.

The second part of the dataset is an inspection record of 144 selected sensors in Taichung City, the largest city in central Taiwan. The inspections were conducted over half a year (from May 29, 2018 to Dec 7, 2018). The malfunctioning sensors are distributed over all areas. In other words, it is inefficient to discover malfunctioning sensors if we simply conduct inspections area-by-area. Out of the 144 randomly selected sensor stations, 28 were malfunctioning. Therefore, the hit rate to discover a malfunctioning sensor based on random selection is approximately 19.4%.

B. Data Preprocessing

We use the monitored PM2.5 values in 2018 as the experimental dataset because the inspection was conducted from May to December 2018.

We found some data quality issues in the dataset of our collected PM2.5 values. First, some values are missing. Second, occasionally, a sensor may transmit multiple different values within a minute. For the first issue, if the missing period is short, we fill in the values with the latest value before the missing period starts. If the missing period is longer than one day, we fill in the missing value by copying the value from the same minute on the previous day. For the second issue, we average the values that were sent within the same minute.

To test the effectiveness of the extra features (humidity and temperature), we also try including these features to predict the PM2.5 values of the target sensor. Referring to Figure 2, we added the target sensor's humidity and temperature measurements at the end of each feature vector.

C. The Sensor Value Prediction (Stage 1)

This section gives the results for Stage 1: predicting the PM2.5 values of the sensors. Table I shows the results of various prediction models. For each model, we show (1) the average R^2 score on the normal sensors (with and without extra features), (2) the average R^2 score on the malfunctioning sensors (with and without extra features), and (3) the difference between (1) and (2) (with and without features).

We have the following observations. First, all the models better predict the future PM2.5 values for the normal sensors than for the malfunctioning sensors. This result demonstrates that our two-stage strategy – predicting sensor values first and ranking sensors by R^2 score from the smallest to the largest to detect the malfunctioning sensors – is likely a reasonable approach. Second, the LSTM model performs best for both the normal sensors and the malfunctioning sensors. However, this is probably not good news since, ideally, we hope a model makes good predictions on the PM2.5 values for the normal sensors and bad forecasts for the malfunctioning sensors. Third, although the overall performance of the ridge regression (labeled as RidgeReg) is not good (second-worst for normal sensors and worst for malfunctioning sensors), the difference of the average R^2 scores between normal sensors and the malfunctioning sensors is the greatest. Consequently, based on our two-stage strategy, ridge regression is likely a favorable model to discover malfunctioning sensors. Finally, including the extra features is slightly helpful since the R^2 score difference between a normal sensor and a malfunctioning sensor generally increases marginally.

D. The Malfunctioning Sensor Prediction (Stage 2)

This section shows the experimental result of Stage 2: predicting the malfunctioning sensors.

We compare the proposed approach with two baseline methods – random inspection and ADF-5 [31]. Specifically, as the current inspection mechanism requires looking through the sampled sensors one by one, this strategy is equivalent to the random inspection baseline method here. The other baseline method, ADF-5, is a famous framework used to detect anomaly sensors for large-scale PM2.5 sensing systems. We use the parameters introduced in the original paper [31], which utilizes 5 neighboring sensors to decide the status (anomaly or normal) of a target sensor.

We report the area under the ROC curve (AUROC), precision at-k ($P@k$), and recall at-k ($R@k$) scores of various models. A brief introduction to these metrics and the reasons for choosing them are discussed below.

The ROC curve shows the relationship between the true positive rate (TPR) and the false positive rate (FPR) using different discrimination thresholds. Since we are dealing with a binary classification problem whose positive ratio (i.e., the number of positive instances divides the total instances) may vary over time, the AUROC score is a proper choice because the ROC curve remains unchanged regardless of the positive ratio and the baseline probability [41].

Table II gives the AUROC scores for various methods. As the machine learning algorithms are nondeterministic by

Using extra features?	Normal Sensors (1)	Malfunctioning Sensors (2)	(1) − (2)	Normal Sensors (3)	Malfunctioning Sensors (4)	(3) − (4)
	False			True		
RidgeReg	0.8215	**0.6005**	**0.2210**	0.8344	**0.6098**	**0.2246**
Lasso	0.8335	0.6778	0.1557	0.8519	0.6700	0.1819
Random Forest	0.7574	0.6156	0.1418	0.7774	0.6319	0.1455
DNN	0.8320	0.6939	0.1381	0.8537	0.7284	0.1253
LSTM	**0.8467**	0.7002	0.1465	**0.8830**	0.7330	0.1500

TABLE I

THE R^2 SCORES OF VARIOUS METHODS FOR THE SENSOR VALUE PREDICTION TASK. WE HIGHLIGHT THE WINNER OF EACH COLUMN IN BOLDFACE. WE HOPE THE MODELS PERFORM POORLY ON THE MALFUNCTIONING SENSORS, SO A LOWER VALUE IN THE "MALFUNCTIONING SENSORS" COLUMN INDICATES A BETTER RESULT. EXTRA FEATURES REFER TO THE HUMIDITY AND TEMPERATURE MEASURES OF THE TARGET SENSOR.

Method	AUROC score
Random Inspection	0.5 (expected value)
ADF-5 [7]	0.624
RidgeReg + MSD	0.7085 ± 0.0135
Lasso + MSD	0.7000 ± 0.0157
Random Forest + MSD	0.6878 ± 0.0063
DNN + MSD	0.6940 ± 0.0072
LSTM + MSD	$\mathbf{0.7090 \pm 0.0072}$

TABLE II

THE AREA-UNDER-ROC CURVE SCORE OF VARIOUS METHODS ON THE TASK OF MALFUNCTIONING SENSOR DETECTION (MEAN ± STANDARD DEVIATION). WE HIGHLIGHT THE WINNER IN BOLDFACE.

Model	$P@10$	$P@20$	$P@30$	$P@40$	$P@50$
Random Inspection	0.194	0.194	0.194	0.194	0.194
ADF-5 [7]	0.300	0.350	0.270	0.330	0.320
RidgeReg + MSD	**0.600**	**0.433**	0.395	**0.375**	**0.337**
Lasso + MSD	0.580	0.430	0.394	0.370	0.320
Random Forest + MSD	0.380	0.370	**0.400**	0.342	0.320
DNN + MSD	0.500	0.430	0.374	0.344	0.312
LSTM + MSD	**0.600**	0.410	0.368	0.332	0.336

TABLE III

THE PRECISION@k ($k = 10, 20, 30, 40,$ AND 50) FOR DIFFERENT METHODS ON THE TASK OF MALFUNCTIONING SENSOR DETECTION. WE HIGHLIGHT THE WINNER OF EACH k BY BOLDFACE.

nature, we report the average and the standard deviation of 5 trials for each machine learning model. Our two-stage strategy with any machine learning model outperforms the current inspection strategy (random inspection) and ADF-5. Among the various machine learning models, LSTM performs the best, followed by ridge regression and Lasso.

We report another metric – precision-at-k ($P@k$). When we can only afford to inspect k sensors, precision-at-k measures the ratio of the discovered malfunctioning sensors among all the inspected sensors. Precision-at-k is defined by Equation 3.

Model	$R@10$	$R@20$	$R@30$	$R@40$	$R@50$
Random Inspection	0.069	0.139	0.208	0.278	0.347
ADF-5 [7]	0.110	0.250	0.290	0.460	0.570
RidgeReg + MSD	0.210	**0.308**	0.423	**0.533**	**0.603**
Lasso + MSD	0.204	0.306	0.422	0.524	0.570
Random Forest + MSD	0.136	0.266	**0.428**	0.484	0.570
DNN + MSD	0.180	**0.308**	0.398	0.484	0.560
LSTM + MSD	**0.214**	0.293	0.394	0.474	0.600

TABLE IV

THE RECALL@k ($k = 10, 20, 30, 40,$ AND 50) FOR DIFFERENT METHODS ON THE TASK OF MALFUNCTIONING SENSOR DETECTION. WE HIGHLIGHT THE WINNER OF EACH k BY BOLDFACE.

$$P@k = \frac{f}{k}, \tag{3}$$

where f is the number of correctly identified malfunctioning sensors when inspecting k sensors.

Table III gives the result of $P@k$ for $k = 10, 20, 30, 40$ and 50. As the inspection in 2018 shows that 28 out of the 144 sampled sensors are malfunctioning, the expected value of $P@k$ for random inspection is $28/144 \approx 0.194$ for all ks. When $k = 10$, using the LSTM model or ridge regression can be three times more effective than random inspection and twice as effective as ADF-5. As k increases, the difference decreases because the LSTM model and the ridge regression are forced to return more sensors whose status may be uncertain.

We also report recall-at-k, which measures the percentage of the discovered malfunctioning sensors among all the sensors that are indeed failed, given that we can only inspect k sensors. The definition of recall-at-k is given by Equation 4.

$$R@k = \frac{f}{N}, \tag{4}$$

where N is the total number of failed sensors, i.e., 28 in our experiment.

Table IV shows the results of $R@k$ ($k = 10, 20, 30, 40, 50$) on different models. When k equals 10, LSTM can be three times and two times more effective than random inspection and ADF-5, respectively. When k becomes larger, ridge regression seems to be the best choice among the compared models.

When comparing the performances of various models on Stage 1 (predicting PM2.5 values) and Stage 2 (predicting the malfunctioning sensors) tasks, the ridge regression model is unimpressive in predicting future PM2.5 values (second-worst according to Table I), but ridge regression is one of the two best algorithms to predict malfunctioning sensors, as demonstrated by Table II, Table III, and Table IV. These results may appear contradictory at first glance. However, as long as a model better predicts the PM2.5 values for the normal sensors than that of the malfunctioning sensors (including the cases where a model makes bad predictions on the PM2.5 values for the normal sensors but makes much worse predictions on that of the malfunctioning sensors), the model can still help us discover the malfunctioning sensors.

V. DISCUSSION

Discovering the malfunctioning sensors in a sensor network is crucial. Unfortunately, effectively finding malfunctioning

sensors based on inspection is challenging, especially when the sensors locate in many different places far away. This paper proposes a two-stage methodology to discover malfunctioning sensors automatically. Experimental results on a nationwide sensor network show that the methodology can effectively identify malfunctioning sensors, suggesting the two-stage method is likely general enough to be applied to other types of sensor networks.

The success of this project requires many factors to coexist. First, the Environmental Protection Agency of Taiwan deploys many air quality sensors and releases the current and historical monitored values. Consequently, it is easy to collect and organize the required PM2.5 information. Second, although the $10,000+$ sensors are distributed widely in Taiwan (whose area is $36,000+$ km^2), Taiwan has extremely convenient road networks, by which one could reach almost anywhere within a couple of hours. Consequently, it is still manageable to dispatch inspectors to most sensor deployment positions from a centralized managing office. Finally, Taiwan's minimum wage[3] is low when compared with countries or economies with similar GDPs per capita.[4] As a result, it is still affordable to hire a maintenance team to conduct regular inspections. The ground truth labels used in our study were directly obtained from the maintenance office. The coexistence of all these factors makes our project a unique case that can effectively collect the monitored values and the status (normal or malfunctioning) of the sensors on a large scale.

A potential problem of the two-stage model is the misidentification of sensor failure and sudden industrial emissions. In both cases, the predicted and monitored PM2.5 values may vary dramatically, so that the malfunctioning sensor detector may regard the corresponding sensors as problematic sensors. A simple workaround for this issue is considering both the R^2 score and the sign of the monitored value subtracted from the predicted value. If the R^2 score is large and the sign is positive, there is probably a sudden emission from the plants. On the other hand, if the R^2 score is high but the sign is negative, indicating that the monitored value is much smaller than the predicted value, the sensor is likely malfunctioning.

We would like to continue refining the model from two directions. First, we would like to integrate more features into the framework. For example, the age of a sensor could be an important factor to indicate its health status. Second, we would like to apply more advanced supervised models for the first stage. In particular, we are most interested in using graphic neural networks since sensors can form a network based on their geographical information. This type of model may be able to discover more problematic sensors with fewer trials.

Another possible future pursuit is automatic sensor calibration. Automatic calibration may look similar to the task in Stage 1, but automatic calibration requires much accurate predictions, which is still beyond the capability of our models in Stage 1. If we could automatically determine the relationship between the actual PM2.5 values and the monitored PM2.5 values of a malfunctioning sensor, the maintenance teams would only be needed when it is difficult to calibrate the sensor automatically. Consequently, we could further diminish the number of required maintenance staff members.

[3]In Jan. 2021, the minimum monthly wage in Taiwan is $\$24,000$ NTD (approximately $\$854$ USD).

[4]Taiwan's GDP per capita is estimated $\$32,123$ USD (nominal) and $\$59,398$ (PPP) in 2021, according to `https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)_per_capita` and `https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(PPP)_per_capita`.

REFERENCES

[1] J. K. Hart and K. Martinez, "Environmental sensor networks: A revolution in the earth system science?" *Earth-Science Reviews*, vol. 78, no. 3-4, pp. 177–191, 2006.

[2] A. Tiwari, P. Ballal, and F. L. Lewis, "Energy-efficient wireless sensor network design and implementation for condition-based maintenance," *ACM Transactions on Sensor Networks (TOSN)*, vol. 3, no. 1, pp. 1–es, 2007.

[3] T. O'Donovan, J. O'Donoghue, C. Sreenan, D. Sammon, P. O'Reilly, and K. A. O'Connor, "A context aware wireless body area network (ban)," in *2009 3rd International Conference on Pervasive Computing Technologies for Healthcare*. IEEE, 2009, pp. 1–8.

[4] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Computer networks*, vol. 38, no. 4, pp. 393–422, 2002.

[5] D. Thomas, M. Orgun, M. Hitchens, R. Shankaran, S. C. Mukhopadhyay, and W. Ni, "A graph-based fault-tolerant approach to modeling QoS for IoT-based surveillance applications," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3587–3604, 2020.

[6] A. L. Clements, W. G. Griswold, A. Rs, J. E. Johnston, M. M. Herting, J. Thorson, A. Collier-Oxandale, and M. Hannigan, "Low-cost air quality monitoring tools: from research to practice (a workshop summary)," *Sensors*, vol. 17, no. 11, p. 2478, 2017.

[7] L.-J. Chen, Y.-H. Ho, H.-C. Lee, H.-C. Wu, H.-M. Liu, H.-H. Hsieh, Y.-T. Huang, and S.-C. C. Lung, "An open framework for participatory pm2. 5 monitoring in smart cities," *IEEE Access*, vol. 5, pp. 14 441–14 454, 2017.

[8] M. A. Zaidan, N. H. Motlagh, P. L. Fung, D. Lu, H. Timonen, J. Kuula, J. V. Niemi, S. Tarkoma, T. Petäjä, M. Kulmala *et al.*, "Intelligent calibration and virtual sensing for integrated low-cost air quality sensors," *IEEE Sensors Journal*, vol. 20, no. 22, pp. 13 638–13 652, 2020.

[9] S. Masoud, N. Mariscal, Y. Huang, and M. Zhu, "A sensor-based data driven framework to investigate pm2. 5 in the greater detroit area," *IEEE Sensors Journal*, 2021.

[10] L. Spinelle, M. Gerboles, M. G. Villani, M. Aleixandre, and F. Bonavitacola, "Field calibration of a cluster of low-cost available sensors for air quality monitoring. part a: Ozone and nitrogen dioxide," *Sensors and Actuators B: Chemical*, vol. 215, pp. 249–257, Aug. 2015.

[11] A. Bigi, M. Mueller, S. K. Grange, G. Ghermandi, and C. Hueglin, "Performance of NO, NO2 low cost sensors and three calibration approaches within a real world application," *Atmospheric Measurement Techniques*, 2018.

[12] E. Esposito, S. De Vito, M. Salvato, V. Bright, R. L. Jones, and O. Popoola, "Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems," *Sensors and Actuators B: Chemical*, vol. 231, pp. 701–713, Aug. 2016.

[13] T. Zheng, M. H. Bergin, R. Sutaria, and others, "Gaussian process regression model for dynamically calibrating and surveilling a wireless low-cost particulate matter sensor network in delhi," *Atmospheric*, 2019.

[14] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1.

[15] A. Appice, A. Ciampi, D. Malerba, and P. Guccione, "Using trend clusters for spatiotemporal interpolation of missing data in a sensor network," *Journal of Spatial Information Science*, vol. 2013, no. 6, pp. 119–153, 2013.

[16] R. Fablet, P. H. Viet, and R. Lguensat, "Data-driven models for the spatio-temporal interpolation of satellite-derived sst fields," *IEEE Transactions on Computational Imaging*, vol. 3, no. 4, pp. 647–657, 2017.

[17] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[18] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, May 2011.

[19] J. Ahn, D. Shin, K. Kim, and J. Yang, "Indoor air quality analysis using deep learning with sensor data," *Sensors*, vol. 17, no. 11, p. 2476, 2017.

[20] X. Luo and J. Yang, "A survey on pollution monitoring using sensor networks in environment protection," *Journal of Sensors*, vol. 2019, 2019.

[21] C.-Y. Lo, W.-H. Huang, M.-F. Ho, M.-T. Sun, L.-J. Chen, K. Sakai, and W.-S. Ku, "Recurrent learning on $pm_{2.5}$ prediction based on clustered airbox dataset," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[22] N. Shaadan, A. A. Jemain, M. T. Latif, and S. M. Deni, "Anomaly detection and assessment of pm10 functional data at several locations in the klang valley, malaysia," *Atmospheric Pollution Research*, vol. 6, no. 2, pp. 365–375, 2015.

[23] L. Lyu, J. Jin, S. Rajasegarar, X. He, and M. Palaniswami, "Fog-empowered anomaly detection in IoT using hyperellipsoidal clustering," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1174–1184, 2017.

[24] P. Ferrer-Cid, J. M. Barcelo-Ordinas, J. Garcia-Vidal, A. Ripoll, and M. Viana, "A comparative study of calibration methods for low-cost ozone sensors in IoT platforms," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 9563–9571, 2019.

[25] A. A. Cook, G. Mısırlı, and Z. Fan, "Anomaly detection for IoT time-series data: A survey," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6481–6494, 2019.

[26] A. Marchioni, M. Mangia, F. Pareschi, R. Rovatti, and G. Setti, "Subspace energy monitoring for anomaly detection @sensor or @edge," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7575–7589, 2020.

[27] P. Ferrer-Cid, J. M. Barcelo-Ordinas, J. Garcia-Vidal, A. Ripoll, and M. Viana, "Multisensor data fusion calibration in IoT air pollution platforms," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3124–3132, 2020.

[28] J. Gu, C. Liu, Y. Zhuang, X. Du, F. Zhuang, H. Ying, Y. Zhao, and M. Guizani, "Dynamic measurement and data calibration for aerial mobile IoT," *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 5210–5219, 2020.

[29] W. Zhang, Q. Lu, Q. Yu, Z. Li, Y. Liu, S. K. Lo, S. Chen, X. Xu, and L. Zhu, "Blockchain-based federated learning for device failure detection in industrial IoT," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5926–5937, 2020.

[30] Y.-W. Lin, Y.-B. Lin, and H.-N. Hung, "CalibrationTalk: A farming sensor failure detection and calibration technique," *IEEE Internet of Things Journal*, vol. 8, no. 8, pp. 6893–6903, 2020.

[31] L.-J. Chen, Y.-H. Ho, H.-H. Hsieh, S.-T. Huang, H.-C. Lee, and S. Mahajan, "Adf: An anomaly detection framework for large-scale pm2. 5 sensing systems," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 559–570, 2017.

[32] A. Fawzy, H. M. Mokhtar, and O. Hegazy, "Outliers detection and classification in wireless sensor networks," *Egyptian Informatics Journal*, vol. 14, no. 2, pp. 157–164, 2013.

[33] L. Yang, Y. Lu, S. X. Yang, Y. Zhong, T. Guo, and Z. Liang, "An evolutionary game-based secure clustering protocol with fuzzy trust evaluation and outlier detection for wireless sensor networks," *IEEE Sensors Journal*, vol. 21, no. 12, pp. 13 935–13 947, 2021.

[34] V. Chatzigiannakis and S. Papavassiliou, "Diagnosing anomalies and identifying faulty nodes in sensor networks," *IEEE Sensors Journal*, vol. 7, no. 5, pp. 637–645, 2007.

[35] O. Ghorbel, W. Ayedi, H. Snoussi, and M. Abid, "Fast and efficient outlier detection method in wireless sensor networks," *IEEE Sensors Journal*, vol. 15, no. 6, pp. 3403–3411, 2015.

[36] R. Wang, Q. Li, H. Yu, Z. Chen, Y. Zhang, L. Zhang, H. Cui, and K. Zhang, "A category-based calibration approach with fault tolerance for air monitoring sensors," *IEEE Sensors Journal*, vol. 20, no. 18, pp. 10 756–10 765, 2020.

[37] A. Ayadi, O. Ghorbel, A. M. Obeid, and M. Abid, "Outlier detection approaches for wireless sensor networks: A survey," *Computer Networks*, vol. 129, pp. 319–333, 2017.

[38] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 12, no. 2, pp. 159–170, 2010.

[39] S. K. Jha, M. Kumar, V. Arora, S. N. Tripathi, V. M. Motghare, A. Shingare, K. A. Rajput, and S. Kamble, "Domain adaptation based deep calibration of low-cost pm2. 5 sensors," *IEEE Sensors Journal*, 2021.

[40] World Health Organization, "WHO air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: global update 2005: summary of risk assessment," Geneva: World Health Organization, Tech. Rep., 2006.

[41] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

Tzu-Heng Lin 's research interests include machine learning, deep learning, and data science. He obtained a Master's degree from the Department of Computer Science and Information Engineering in National Central Unversity in 2021. Tzu-Heng's GitHub: `https://github.com/NeilLin1117`.

Xin-Ru Zhang obtained a Master's degree and a Bachelor's degree from the Department of Computer Science and Information Engineering at National Central University in 2021 and 2019, respectively. She is interested in various machine learning topics such as deep learning and data analysis.

Chia-Pan Chen is an Environmental Management Senior Engineer at the Industrial Technology Research Institute. He is the Project Manager for the Taiwan Air Wide Array Network project, mainly in controlling and monitoring the progress of sensor deployment nationwide, data quality testing and verification and checking with third party. He is also the Principal Investigator for National Industrial Park Mutual Aid Plan, mainly in preventing disaster and assisting with chemical hazard emergency response and rescue. He received a Master's degree and Bachelor's degree from the Institute of Civil and Hydraulic Engineering, Feng Chia University and Department of Civil Engineering, Feng Chia University, respectively.

Jia-Huei Chen is an Associate Research Fellow at the Industrial Technology Research Institute and an Associate Researcher for the Taiwan Air Wide Array Network project, in which she applies spatial statistics, machine learning, and IoT technology to assist in air quality data analysis. She is a Ph.D. Candidate at National Taiwan Normal University. She was a GIS analyst for an environmental consulting company, mainly in designing the geography information system, analyzing spatial data and managing project progress. She obtained a Master's degree and Bachelor's degree both from the Department of Geography, National Taiwan Normal University.

Hung-Hsuan Chen is an Associate Professor at the Department of Computer Science and Information Engineering at National Central University (NCU). He is interested in data-related research topics such as machine learning, information retrieval, and graph analysis. He is also interested in applying these techniques to various application domains such as AIoT, recommender systems, and social networks. He was a researcher at the Industrial Technology Research Institute. He obtained his Ph.D. degree from the Pennsylvania State University in 2013.