IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 29, 2020, accepted April 27, 2020, date of publication May 11, 2020, date of current version May 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2993547

Real Time Localized Air Quality Monitoring and Prediction Through Mobile and Fixed IoT Sensing Network

DAN ZHANG[1,2] AND SIMON S. WOO[3,4]

[1]Department of Computer Science, The State University of New York (SUNY) Korea, Incheon 21985, South Korea
[2]Department of Computer Science, Stony Brook University, Stony Brook, NY 11794, USA
[3]Computer Science and Engineering Department, Sungkyunkwan University, Suwon 16419, South Korea
[4]Department of Applied Data Science, Sungkyunkwan University, Suwon 16419, South Korea

Corresponding author: Simon S. Woo (swoo@g.skku.edu)

This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, through the ICT Consilience Creative Program supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP) under Grant IITP-2020-2011-1-00783, in part by the Energy Cloud Research and Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT under Grant 2019M3F2A1072217, in part by the Basic Science Research Program through the NRF of Korea funded by the Ministry of Science and ICT (MSIT) under Grant 2017R1C1B5076474 and Grant 2020R1C1C1006004, and in part by the Institute for Information and Communication Technology Promotion (IITP) funded by the Korea Government (MSIT) (Regional strategic industry convergence security core talent training business) under Grant 2019-0-01343.

ABSTRACT Air pollution and its harm to human health has become a serious problem in many cities around the world. In recent years, research interests in measuring and predicting the quality of air around people has spiked. Since the Internet of Things (IoT) has been widely used in different domains to improve the quality life for people by connecting multiple sensors in different places, it also makes the air pollution monitoring more easier than before. Traditional way of using fixed sensors cannot effectively provide a comprehensive view of air pollution in people's immediate surroundings, since the closest sensors can be possibly miles away. Our research focuses on modeling the air quality pattern in a given region by adopting both fixed and moving IoT sensors, which are placed on vehicles patrolling around the region. With our approach, a full spectrum of how air quality varies in nearby regions can be analyzed. We demonstrate the feasibility of our approach in effectively measuring and predicting air quality using different machine learning algorithms with real world data. Our evaluation shows a promising result for effective air quality monitoring and prediction for a smart city application.

INDEX TERMS Time-series prediction, air quality measurement, machine learning.

I. INTRODUCTION

Due to rapid urbanization and industrialization, many countries around the world are facing a critical crisis of air pollution. Air pollution has become a threat to public health and a heavy influential factor on citizen's daily activity. In metropolitan cities in developing countries bothered by problems of air pollution, such as Beijing and Delhi, people usually need to wear a mask before going out [1]. Besides, outdoor activities are also constrained by the intra-day air quality.

Air pollution is caused by the presence of different air pollutants. The primary air pollutant gases are nitrogen dioxide (NO_2), carbon monoxide (CO), ozone (O_3) and sulphur dioxide (SO_2) [2]. Another type of air pollutants is air particulate matter (PM). Among them, $PM_{2.5}$ and PM_{10} are of particular concerns to people, which refers to atmospheric particulate matter that have a diameter of less than 2.5 μm and 10 μm. These particles can cause many respiratory or cardiovascular diseases [3]. Thus, many cities have built their own air quality monitoring stations and publish the real-time air quality information every hour. As the concern for air pollution increases, its becoming increasingly critical to measure the air quality around people [4], [5], which inform people about when is safe to perform outside activities and help them plan better routes to reach their destinations. Typically, monitoring stations at fixed locations is the conventional approach for atmospheric factor monitoring for a large geographical district.

The associate editor coordinating the review of this manuscript and approving it for publication was Junaid Arshad.

While it is not difficult to implement such fixed sensor based monitoring system, it faces several challenges. First, huge investment is involved in building and deploying monitoring units to cover a large area. Also, it is highly dependent on neighboring environments and tends to be less accurate for farther areas. In areas close to the roads, even small distances can make a huge difference in air quality data measurement from car pollutions. Hence, new ways to collect air quality information in a cheaper and more flexible way and provide detailed air quality prediction is in demand.

To address these issues, one possible solution is to make the sensors mobile using Internet-of-Things(IoT). For example, attaching sensors on moving cars or drones proved to be a feasible method [6]. In this work, we developed the IoT devices to monitor air quality. We collected air pollution data by mounting a sensor to a car and moved around the city of Incheon, Republic of Korea. This data is then preprocessed and stored in our server. One major advantage of using a mobile sensor is that it provides the very first hand air pollution information for an area at a particular time, when the car was moving through there. we can also cover more geographical regions and have more accurate localized information with mobile IoT sensors. While a static fixed sensor can provide continuous feed of information about a particular area, it is not easy with a mobile sensor. However, this can be minimized by having multiple mobile sensors or assigning smaller coverage area to a mobile sensor.

In this work, we propose a hybrid approach, where we deploy multiple static sensors as well as IoT mobile sensors to effectively monitor air quality. The static sensors can provide a holistic view by providing a continuous feed of information. On the other hand, mobile sensors can provide more accurate data about specific areas to reduce the error from static sensors. In this paper, we build a prediction model to utilize the collected data and provide rapid information about the air quality around people. We also developed a visualization tool to better analyze and forecast air quality and provide insights to both professional researchers and ordinary users. The main contributions of our work are summarized as follows:

- We proposed a hybrid approach to integrate fixed and mobile IoT sensors to measure and predict air quality data.
- We demonstrated the feasibility and effectiveness of our approach by analysing the prediction result with different machine models.
- We developed a visualization tool to show the relative distribution of the air pollutants with a focus on PM_{10} and $PM_{2.5}$, where it provides an intuitive understanding of the air quality around people.

The rest of our paper is organized as follows: Section 2. presents the related work on different air quality measurement and prediction methods. Section 3 describes the development of IoT sensors and data processing. Section 4 explains our models and algorithms. The experimental setup and results are reported in Section 5, and an analysis of the results is provided in Section 6. We summarize our work and offer conclusion in Section 7 and Section 8.

II. RELATED WORK

To measure the air quality, several monitoring methods have been proposed and utilized. In Zheng *et al.*'s research [7], they use public and private web services as well as a list of public websites to provide real-time meteorological, weather forecasts and air quality data for their forecasting. Small unmanned aerial vehicles are used in the work of Alvarado *et al.* [8] as a methodology to monitor PM_{10} dust particles, where they can calculate the emission rate of a source. With the development of smart city technologies, IoT devices have been shown to be an effective option to collect real time weather, road traffic, pollution and traffic information. Thus, IoT devices are also considered to enable air quality analysis [9].

In addition to the fixed sensors, public transportation infrastructure such as buses has been used to collect air quality data [10]. Also, there is one project [11] engaged the entire community members in collecting data and developed an online air quality monitoring system based on it, which is also called crowdsourcing. Hasenfratz *et al.* [12] utilized sensor nodes to build a thousand models targeting at different time periods. All these aforementioned methods are either costly or time consuming. In our work, we explore the use of fixed and mobile IoT sensors together to improve the prediction performance, which has not been researched much yet.

To meet the increasing query frequency of air quality in real time and also to enable citizens to react instantly to the pollution, there has been a large body of work on building connected monitoring sensor networks to share the current air quality information with them [13]. Garzon et.al presented in [14] an air quality alert service. Their service continuously determines the areas, where the level of certain matter concentration exceeds the preset threshold, and notify users if they entered them. Maag *et al.* [15] proposed a multi-pollutant monitoring platform using wearable low-cost sensors. Compared with above methods, our system can serve the similar functions to end users practically with either fewer sensors or less demand for computation.

For prediction, regression models are commonly used in the area of air quality prediction. A multivariate linear regression model for predicting $PM2.5$ of short-period time is proposed in Zhao's work [16], which includes other gaseous pollutants such as SO_2, NO_2, CO and O_3. As deep learning emerged as an effective method in many applications, time series data of air pollution based on different network models have been also extensively studied and developed. Novel models such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit Network (GRU) have been proved to be a powerful sequential structures in predicting future values of air quality [9], [17] . Yi *et al.* [18] proposed a deep distributed fusion network to learn the characteristics of spatial dispersion and capture all the influential factors that may have a

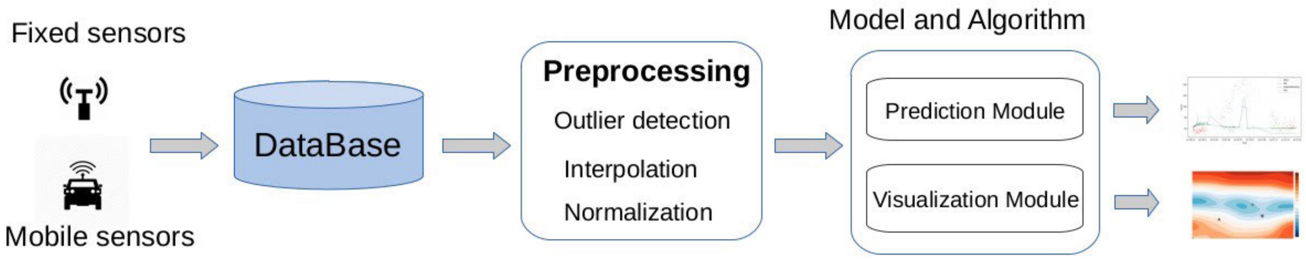


FIGURE 1. The structure of proposed system for monitoring and prediction of air pollution.

direct or indirect effect on air quality. These aforementioned technologies fits non-linear models flexibly but usually being short of offering insight to the hidden mechanism. In addition, they have not shown to necessarily outperform classical regression models in many scenarios [19]. There are also a lot of researches concentrate on approaches to model and simulate the pollutants for prediction [20]. With a small amount of data set oriented in our project, we decided to take conventional regression models as our baseline methods because of computation efficiency, while yielding favorable results.

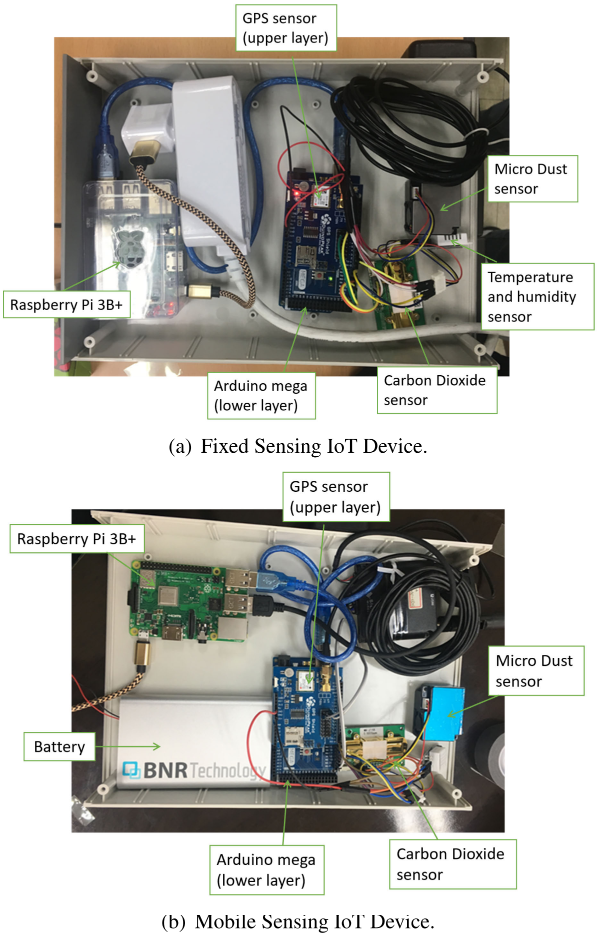
III. IMPLEMENTATION

In this section, we first describe the design and implementation of IoT sensor device deployed in our research. Our deployment and data collection are performed in Songdo [21], South Korea, which is envisioned to be developed as a smart city. Next, we explain the preliminary processing of the acquired raw data and describe how we store and transmit the collected and cleaned data. Then, we further present the user interface to check the collected data for our analysis. Figure 1 describes the overall architecture of our proposed system.

A. IoT SENSOR INSTRUMENT DESIGN

We assembled two types of sensing devices from off-the-shelf parts, one for fixed locations and the other type for moving cars. In total, we developed six IoT sensor devices, where three of them are deployed in three different fixed locations and the other three are mounted on data collection cars. The subsystems of the air quality monitoring modules are presented in Fig. 3, and the functions of the sensors are described as following:

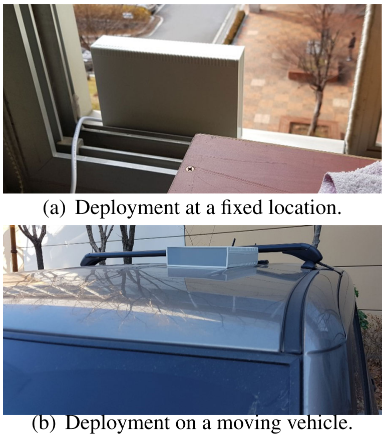
- **Temperature and humidity sensor**: We have a single sensor that can measure both temperature and humidity. The humidity sensor provides an accuracy of 2%, whereas the temperature sensor has an accuracy of $0.5°C$. They have measurement ranges of $0 \sim 100\%$ and $-40 \sim 80°C$, respectively.
- **Micro Dust sensor**: This sensor measures both $PM_{2.5}$ and PM_{10}. The range of these measurements is from $0 \sim 999.9\mu g/m^3$. The Government of Korea considers $PM_{2.5}$ and PM_{10} values of over $35\mu g/m^3$ and $100\mu g/m^3$ averaged through a day to be dangerous for human health. Thus, our micro dust sensor covers the entire range that is relevant for human health.
- **Carbon Dioxide sensor**: Our carbon dioxide sensor can measure CO_2 within a range of $0 \sim 10000ppm$, with an accuracy of $5ppm(0 \sim 2000ppm)$, $10ppm(2000 \sim 5000ppm)$, and $20ppm(5000 \sim 10000ppm)$. Note that since in a natural scenario, the proportion of CO_2 is around 0.03%, this level of accuracy is sufficient for our purpose.
- **Raspberry Pi 3B+**: The Raspberry Pi is connected to LTE using a dongle. Its main function is to process the



(a) Fixed Sensing IoT Device.

(b) Mobile Sensing IoT Device.

FIGURE 2. Two types of air quality monitoring IoT sensor modules.


(a) Deployment at a fixed location.

(b) Deployment on a moving vehicle.

FIGURE 3. Deployment at fixed as well as moving scenario. The first deployment is on the window of a building. The second deployment is on top of a car.

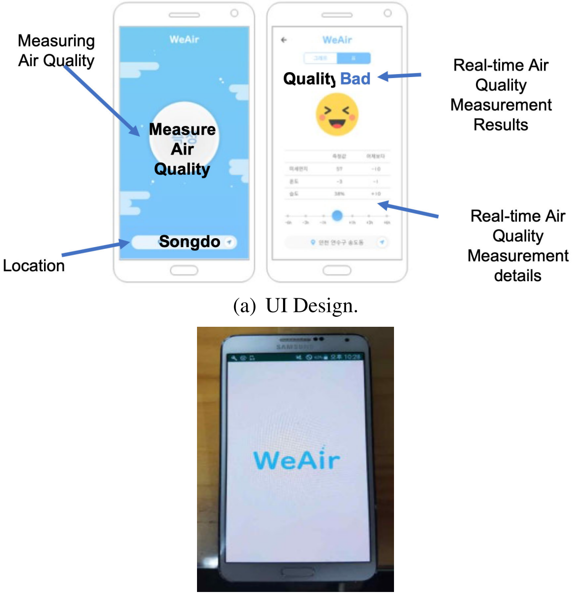
sensor data and send it over the internet to the cloud server.

- **Arduino mega**: This implements the protocol for sending data over the VoLTE network.
- **GPS sensor**: This GPS sensor is connected to the Arduino, and provides an accuracy of close to 1 *m*.
- **Battery**: We use a power bank with a capacity of 7,000 mAh. The overall power consumption of our setup is close to 1A. Thus, our setup can run continuously for around 7 hours without a single charge.

B. SOFTWARE DEVELOPMENT AND PRE-PROCESSING OF ACQUIRED DATA

In this section, we describe the software systems that we designed, which run on top of the IoT sensors and transmit collect data back to our system. Also, we present other software and database components needed in our system to pre-process acquired data.

- **Communication Software:** We constructed a wireless communication and GPS system to transmit acquired data back to databases for analysis. The geo-tagged data which is stored in Raspberry Pi is transmit over Voice over Long-Term Evolution(VoLTE) once per second to our central server in Songdo area.
- **Database**: We design the database to store the collected real time sensor values from fixed as well as mobile IoT sensors. The data fields are: 1) time, 2) GPS location, 3) temperature, 4) humidity, 5) CO_2, 6) PM_{10}, and 7) $PM_{2.5}$, where all the collected values are stored in database as shown in Fig. 5. (a). In the areas with weak GPS signals, such as indoors and tunnels, we approximate the value according to the latest neighboring data. Further, we discard out-of-range data during the pre-processing.
- **Cloud Server and Data Mapping**: We use a cloud server for our system, where the server manages the data and provides an interface for analyzers to check the details of the collected real air pollution value by using a DatePicker Dialog. It allows us to choose sensor number and other fields such as dates. There are two main functions of our web portal: (1) Time flow of the real data of different categories with marked min and max values, and (2) Google Map API is integrated in the website to visualize the traces of the moving cars during the chosen time period as shown in Fig. 5(b). All the cars followed different paths randomly but tried to cover the entire area as much as possible. All the stored data can be downloaded in the form of Excel spreadsheet for later analysis.
- **User Interface (UI):** In addition, we developed the User Interface (UI) App so that users can log in our developed APP using their own account and check the air quality data around them immediately. The example of user interface is provided in Fig. 4, where APP can measure the real time air quality measurements and display those.


Measuring Air Quality
WeAir
Measure Air Quality
Songdo
Location
WeAir
Quality Bad
Real-time Air Quality Measurement Results
Real-time Air Quality Measurement details
WeAir

(a) UI Design.

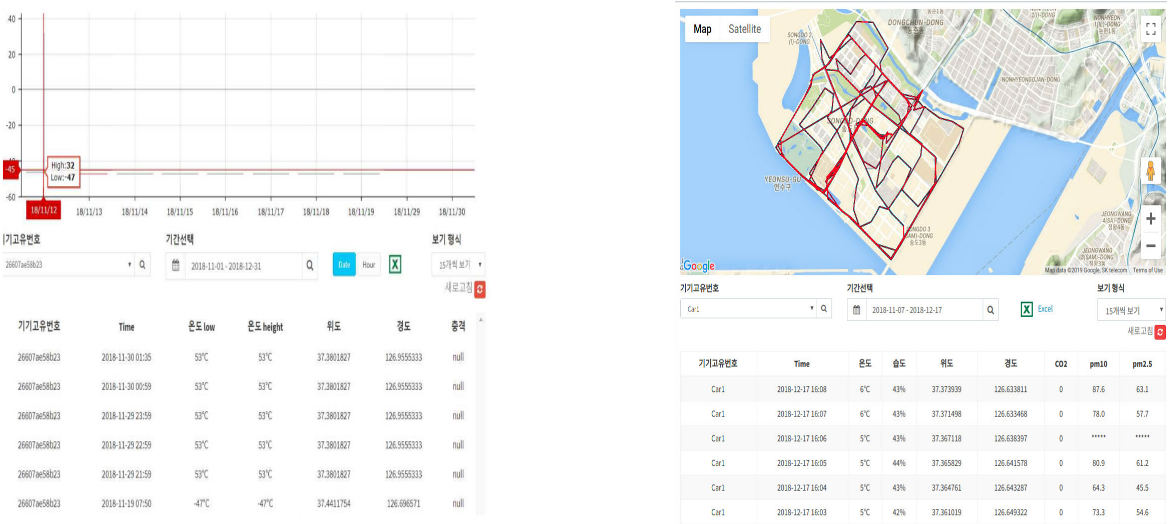
(b) The Real UI interface on phone.

FIGURE 4. User Interface of our developed application WeAir.

C. PRE-PROCESSING OF ACQUIRED DATA

Since the acquired data would contain noise, missing values, etc, we need to pre-process the acquired data to develop a robust prediction model. We employ the following techniques to pre-process data:

- **Outlier detection**: Since sudden changes in the collected data usually means an outlier, we calculated the discrete differences of measured sensor values along the timeline to detect the outliers. That is, measured samples with a discrete difference beyond the interval $[-0.5, 2]$ are removed from our data set.
- **Interpolation**: We choose Gaussian Process Regression (GPR) [22] as our interpolation method because it assists



기기고유번호: 26607ae58b23
기간선택: 2018-11-01 - 2018-12-31
보기 형식: 15개씩 보기
새로고침

기기고유번호	Time	온도 low	온도 height	위도	경도	충격
26607ae58b23	2018-11-30 01:35	53°C	53°C	37.3801827	126.9555333	null
26607ae58b23	2018-11-30 00:59	53°C	53°C	37.3801827	126.9555333	null
26607ae58b23	2018-11-29 23:59	53°C	53°C	37.3801827	126.9555333	null
26607ae58b23	2018-11-29 22:59	53°C	53°C	37.3801827	126.9555333	null
26607ae58b23	2018-11-29 21:59	53°C	53°C	37.3801827	126.9555333	null
26607ae58b23	2018-11-19 07:50	-47°C	-47°C	37.4411754	126.696571	null

(a) Details of the collected data in a selected day, which shows the exact value of time, temperature, humidity, CO_2, PM_{10}, $PM_{2.5}$, latitude and longitude.

기기고유번호: Car1
기간선택: 2018-11-07 - 2018-12-17
보기 형식: 15개씩 보기
새로고침

기기고유번호	Time	온도	습도	위도	경도	CO2	pm10	pm2.5
Car1	2018-12-17 16:08	6°C	43%	37.373939	126.633811	0	87.6	63.1
Car1	2018-12-17 16:07	6°C	43%	37.371498	126.633468	0	78.0	57.7
Car1	2018-12-17 16:06	5°C	43%	37.367118	126.638397	0	*****	*****
Car1	2018-12-17 16:05	5°C	44%	37.365829	126.641578	0	80.9	61.2
Car1	2018-12-17 16:04	5°C	43%	37.364761	126.643287	0	64.3	45.5
Car1	2018-12-17 16:03	5°C	42%	37.361019	126.649322	0	73.3	54.6

(b) Trace of the selected patrolling car in a selected day, which shows the areas our collected data covers.

FIGURE 5. Interface for analyzers to check the details of the collected air pollution values on cloud server.

in reaching the best prediction accuracy in our experiments, and the effect of different interpolation methods will be discussed in Section 6.

- **Data normalization**: Since data are measured at different scale, we normalize the sensor measurement between 0 and 1 using Eq. 1. Thus, we can use normalized dataset for developing the air quality prediction models:
$$x^* = \frac{x - min}{max - min}, \tag{1}$$
where *max* and *min* are the maximum and minimum value of the whole dataset and x^* is the data value after the normalization.

IV. PREDICTION ALGORITHMS AND MODEL DEVELOPMENT

In this section, we introduce our prediction model and briefly discuss algorithms we used. Since random forest (RF) [23], support vector machine (SVM) [24], and gradient boosting machine (GBM) [25] are commonly recognized as the most powerful algorithms in many machine learning applications [26], [27], we deployed random forest regressor (RFR) [28], support vector regressor (SVR) [29] and gradient boosting regressor (GBR) [30] for predicting air quality. We initially considered these approaches and explain more details in the following sections.

A. SUPPORT VECTOR REGRESSOR (SVR)

The objective of SVR is to determine a hyper-plane in the space generated by mapping training data in its original space to a higher dimensional feature space, and the hyper-plane can minimize the deviation of all sample points from it. Consider the training data set $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$, where $\mathbf{x} \in \mathbb{R}^n$, $y \in \mathbb{R}$ where m corresponds to the number of training data, then the regression problem can be formulated as:

$$\min_{w,b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} l_\epsilon(f(\mathbf{x}_i) - y_i). \tag{2}$$

Here C is a constant, $f(\mathbf{x})$ is the hyper-plane represented as $f(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x}+b$, and l_ϵ is the cost function which is minimized in Eq. 2:

$$l_\epsilon(z) = \begin{cases} 0 & \text{if } |z| \leq \epsilon, \\ |z| - \epsilon & \text{otherwise,} \end{cases} \tag{3}$$

where ϵ is the deviation which we can bear with at the most. Basically, the equations build a interval-zone with the width of 2ϵ centered on $f(\mathbf{x})$. In our research, feature vector $\mathbf{x}$ consists of the properties of time, longitude, and latitude information collected from sensors, and y represents a collected value from air pollutants set CO_2, $PM_{2.5}$ and PM_{10}.

B. RANDOM FOREST REGRESSOR (RFR)

RFR is fast in learning, and is capable of handling a large number of input variables yet yielding high accuracy. RFR randomly draws samples from the original dataset with replacement, which is also called bootstrap, and grows an unpruned regression tree for each of the samples, then average the unweighted outputs of multiple decision trees to obtain the final result as follows:

$$\bar{h}(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^{K} h(\mathbf{x}; \theta_k), \tag{4}$$

where $h(\mathbf{x}; \theta_k)$ is a collection of tree predictors with $k = 1, \ldots K$, θ_k is random vector, which characterizes the kth

TABLE 1. Details of the dataset collected from our three mobile IoT sensor boxes.

Date	Time Span (Car 0)	Time Span (Car 1)	Time Span (Car 2)	No. of instances
2018/12/10	10:21 – 14:40 17:05 – 19:22	/	/	398
2018/12/11	10:12 – 17:24	/	/	432
2018/12/12	10:23 – 18:04	13:48 – 14:37 17:18 – 20:30	14:40 – 23:59	1261
2018/12/13	13:14 – 16:23 22:03 – 22:34	09:21 – 10:37 15:19 – 17:30	00:00 – 23:59	1867
2018/12/14	09:17 – 13:00 14:53 – 15:40	10:25 – 15:28	09:40 – 09:50 15:14 – 15:45	614
2018/12/17	09:43 – 14:39	09:13 – 11:20 14:08 –16:08	09:27 – 18:19	1075
2018/12/18	09:59 – 11:00 16:20 – 16:31	10:34 – 11: 43 14:29 – 17:47	09:58 – 15:47	688
2018/12/19	09:52 – 13:32 15:08 – 15:20	12:02 – 17:36	09:57 – 11:10 14:23 – 18:10	986

RF tree, $\mathbf{x}$ represents the observed input which are assumed to be independently drawn from the joint distribution $(\mathbf{x}, y)$. Similarly, x represents time, longitude, and latitude information collected from sensors, and y represents a collected value from air pollutants set CO_2, $PM_{2.5}$ and PM_{10}.

C. GRADIENT BOOSTING REGRESSOR (GBR)

Gradient descent tries to minimize a function by moving in the opposite direction of the gradient, and it is a fundamental optimization algorithm in the area of machine learning. Boosting is known as an ensemble method that can improve the prediction performance of classification or regression [27]. It constructs additive regression models by iteratively adding basis functions which can further reduce the designed cost function:

$$f(\mathbf{x}) = \sum_{m=1}^{M} \beta_m h(\mathbf{x}; a_m), \tag{5}$$

where the function $h(\mathbf{x}; a_m)$ is the basis function that are usually chosen to be simple representation of $\mathbf{x}$ with parameters $a = \{a_1, a_2, \ldots\}$, and β_m are the expansion coefficients with $m = 1, 2, \ldots, M$. Regression trees are used as a basic function in our model. With our dataset, the features used in $\mathbf{x}$ and y are the same as described in previous models.

V. EXPERIMENT

We have chosen the geographic region of Songdo, Incheon, Korea as a location for conducting our experimental study, where Songdo has been developed as one of the smart cities in South Korea. In the experiment, Songdo region is spatially segmented into 100 zones, 10 × 10 grids as shown in Fig. 6 in the latitude range 126.616° to 126.700° and the longitude range 37.348° to 37.401°, where the red dots represents data collection points by mobile and fixed sensors. With more sensors operating in the future, we can divide the area into more grids which enables a higher resolution service to the public. Three fixed sensors are marked with a yellow star respectively in the map. As we can observe, the density of data collecting points are higher at the fixed sensors' position. In order to cover the entire Songdo area as much as possible, three cars are mounted with our mobile IoT box and navigated the road from Dec. 10th to Dec. 14th, 2018 and from Dec. 17th to Dec. 19th, 2018. Each day all the sensors are calibrated at both pre-deployment and post-deployment stage. The details, such as time intervals and the number of collected data instances are provided in Table 1, and we use the name Car0, Car1 and Car2 to differentiate the three mobile IoT sensors.

A. DATASET

Both the fixed and mobile sensors collect the same format of dataset. The fixed sensors collect air quality data every minute from the three chosen locations in Songdo area shown as yellow stars in Fig. 6. For each fixed sensor, the data collection time periods span all day, basically from morning to night. The mobile sensors, however, collect the air pollution data only a few hours per day, but the whole dataset in general also covers all hours of a day.

The geographical locations of these sensors are presented in Fig. 6, where each icon stands for a sensor. The horizontal and vertical lines of the grids are cut according to latitude and longitude, and spaced evenly to grant same size grids. Each collected data instance consists of the sensor box's longitude and latitude, timestamp, temperature, humidity, and concentration value of CO_2, $PM_{2.5}$ and PM_{10}.

The observed time series data of $PM_{2.5}$ and PM_{10} collected from the moving sensors for the entire region are depicted in Fig. 7. We averaged the data collected from all the moving sensors at each moment. Along the X axis is the timeline and Y axis represents the pollutants' observed value, and the quantity unit for PM_{10} and $PM_{2.5}$ is $\mu g/m^3$.

B. PERFORMANCE METRIC

Based on the previous day's ground truth y_i from mobile sensors, we evaluate the prediction $\hat{y}_i$ and the model's performance according to Root Mean Square Error (RMSE), which

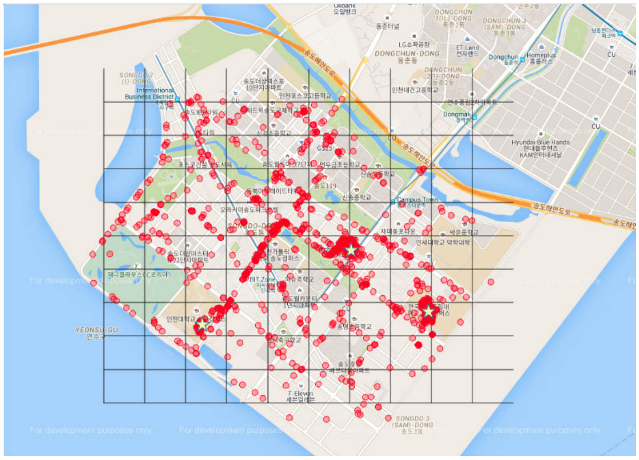


FIGURE 6. Illustration of the geographical grids for Songdo on Google Map, where the red dots are sensor measurement location and three yellow stars indicate the fixed sensor locations.

is adopted as an error criteria and defined by Eq. 6 as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(|y_i - \hat{y}_i|)^2}. \qquad (6)$$

VI. RESULTS

In total, three fixed sensors and three mobile sensors generated 13,128 measurements from Dec. 12th, 2018 to Dec. 19th, 2018. The entire dataset is divided into non-overlapping two parts for training and test, while the time intervals in the training and test datasets varies from task to task.

A. OVERALL PERFORMANCE COMPARISONS WITH DIFFERENT PREDICTION ALGORITHMS

In this section, we used RFR, SVR and GBR to validate the overall performance of our proposed air quality prediction model. We split the entire dataset into 8 non-overlapping training and test pairs, where each individual day from Dec. 12th to Dec. 19th is a test dataset and all the prior date forms the training dataset, respectively. Table 2 presents the overall performance of different regression algorithms across various test days. Values in bold indicates the best prediction in a specific testing day. We can see that in general, GB regressor achieved the highest prediction accuracy as shown in Table 2, while RFR and SVR has marginally better performance in one or two days.

We provided sample prediction results in Fig. 8 across different time periods. A few trends are visible in the results. First, we find that the values of PM_{10} is greater than that of $PM_{2.5}$, as shown in Fig. 7. As expected, there is usually less $PM_{2.5}$ content in the environment for $PM_{2.5}$ than PM_{10}. Second, we find that predictions for the good air quality days are much better than the polluted days. For example, the fine particles' real value in Fig. 8(c) and Fig. 8(d) are much higher than other days. In addition, the RMSE value of the same day, Dec. 16th, is also higher than other days, where they are 21.6 for PM_{10} and 15.8 for $PM_{2.5}$ as shown in Table 2, respectively. Finally, we find that the Gradient Boosting (GB) technique is the most responsive to sudden changes in patterns. While SVR and RFR are effective in finding the overall trends, they do not provide good prediction in the short term.

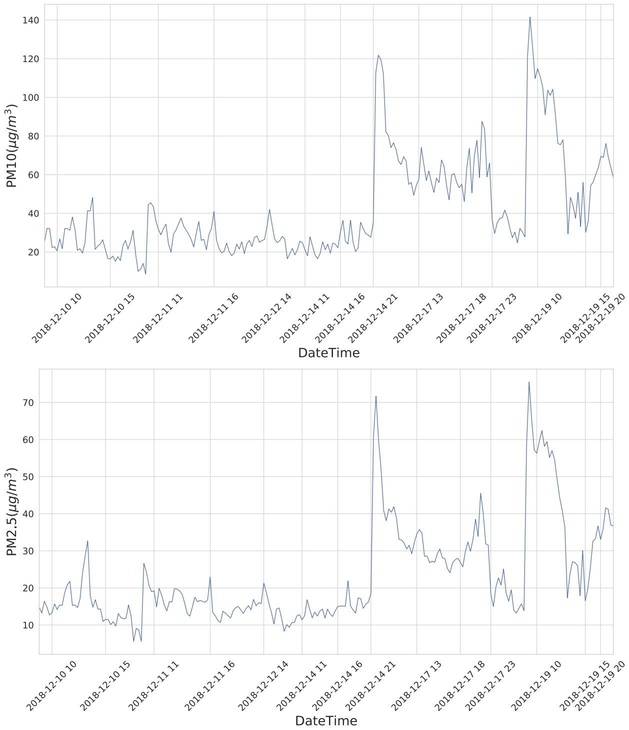


FIGURE 7. Observed real micro dust data from our three mobile sensors at Songdo, Incheon and the values are averaged when there is more than one sensor working at each moment.

B. ACCURACY PERFORMANCE WITH DIFFERENT NUMBER OF GRIDS

For evaluation, we select Dec. 19th, 2018 as the test data and data from Dec. 10th, 2018 to Dec. 18th, 2018 as the training data. Since $PM_{2.5}$ and PM_{10} are our major interest, we focus on the prediction accuracy comparison on $PM_{2.5}$ and PM_{10}, and we chose GBR as our prediction algorithm, as it outperforms the other two methods in the previous evaluation section in general. As shown in Table 3, we counted the number of samples in each of the 100 grids and divide the number of samples into 6 intervals based on its distribution: $0 \sim 10$, $11 \sim 20$, $20 \sim 50$, $50 \sim 100$, $100 \sim 200$ and above 200. Then, we calculated the number of grids in each category and all these grids' RMSE of prediction. At last, we averaged the RMSE of all the grids in that specific category.

We can observe that an increase in the number of training samples in a grid leads to lower RMSE, and thus higher prediction accuracy. It demonstrates the validity of our methods and indicates that air quality prediction can be improved with

TABLE 2. Comparison of the prediction results with different regression algorithms.

	Test date	12.12	12.13	12.14	12.15	12.16	12.17	12.18	12.19
$RMSE(PM_{10})$	RFR	11.2	13.4	7.2	23.16	26.1	19.9	7.9	17.6
	SVR	**9.8**	14.0	7.3	22.7	25.9	22.5	9.5	16.5
	GBR	10.3	**10.4**	**6.6**	**20.3**	**21.6**	**19.2**	**7.1**	**15.2**
$RMSE(PM_{2.5})$	RFR	9.8	11.3	**5.7**	**21.2**	17.5	**15.9**	6.8	9.5
	SVR	9.2	11.6	6.7	24.7	16.1	17.3	8.1	12.5
	GBR	**7.8**	**7.6**	6.2	22.3	**15.8**	16.2	**6.0**	**7.9**

collecting more data in the future. We observed a similar result for carbon dioxide as well.

C. PERFORMANCE WITH DIFFERENT INTERPOLATION METHODS

As discussed before, the collected data is very sparse on the geographical grids in a specific time point and the dispersion characteristics of the fine particles are complex to model. Therefore, different interpolation techniques are examined in our model to fill the missing air pollution data in all the other grids. In order to check whether the interpolation strengthens our prediction, we compare three different interpolation methods with our baseline (no interpolation). Since conventional interpolation method Kriging [4], [31] shares the same mean value and confident interval with Gaussian Process Regression (GPR), we choose linear interpolation and GPR with different kernels (Gaussian and Cauchy) for our investigation. We used the same training and test dataset as described in the previous section.

Table 4 presents the overall prediction results comparing different interpolation methods (Linear interpolation, GPR + Cauchy kernel, and GPR + Gaussian kernel) with the original baseline without interpolation. A clear improvement on the accuracy can be observed as shown in Table 4 across all training time intervals. GPR + Gaussian kernel outperforms both Cauchy kernel and linear interpolation in the final results for PM_{10} and $PM_{2.5}$.

D. PERFORMANCE ON INTEGRATING MOVING IoT SENSORS

To demonstrate the effectiveness of our hybrid approach in air quality prediction, we compared the performance between using 1) fixed sensors only vs. 2) both fixed sensors and mobile sensors (our approach). In this evaluation, we also use GBR as the analysis tool, and tested on 4 different days chosen from the entire dataset. In each test, data collected from the previous two days ahead of the test date is utilized as the training data. We calculated and compared the prediction RMSE from all the grids for both $PM_{2.5}$ and PM_{10} using GBR in all the four test days and averaged it to obtain the final RMSE as shown in Table 5.

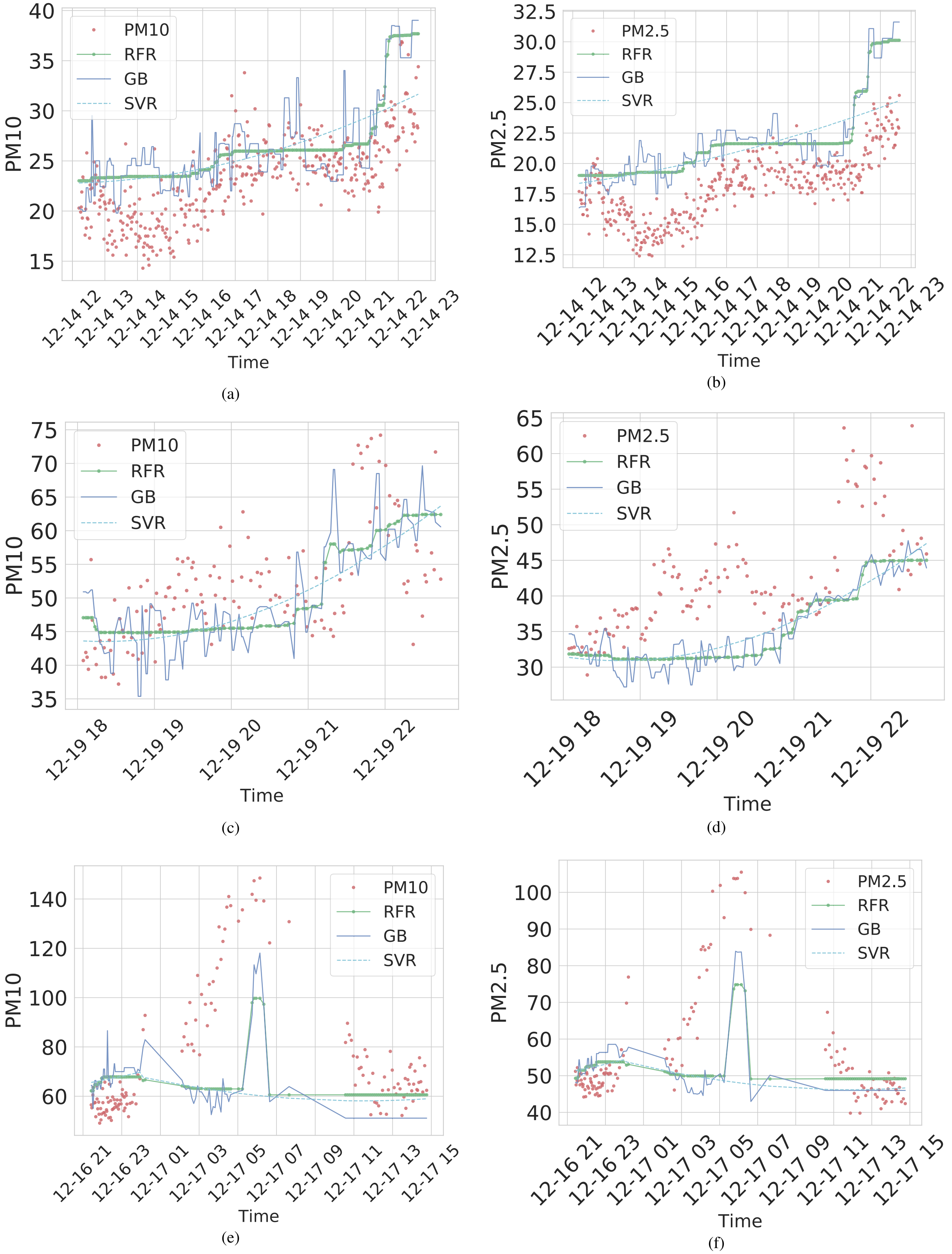
Details of the training and testing set splits and the final results are presented in Table 5, where the prediction with hybrid fixed and mobile sensors outperformed the one with only fixed sensors in all the test days. With the value of hybrid method marked in bold, it is clear to observe that hybrid sensors method can improve the overall prediction accuracy, compare to using only fixed sensors by 7.0% for PM_{10} and 6.5% for $PM_{2.5}$ on average. Thus, our proposed method enhanced the performance of air quality prediction.

E. VISUALIZATION

It is challenging to visualize the air quality data because there are multiple sensors data which are moving around. Common method for visualizing air quality data [32] is to overlay a contour map on the geographical map. The pattern in the contour map is simple, where only limited polluted locations are identified and presented as point sources. In this way, the surrounding area's air quality value is roughly estimated without considering the integrated impact of different pollution sources. We studied the relative distribution of the pollutants in Songdo area and drew a heatmap to visualize the hidden relationship of air quality in the whole area.

The map of our experimental area is shown in Fig. 9, whose shape is very close to a rectangle. Thus, we defined the heatmap as a 1,000 × 1,000 pixel image. Since each pixel in the generated visualization graph corresponds to a geographically position on map in Fig. 9, we assign a color value to each pixel according to the air pollution factor value at that geographical location. This task is implemented in the following three steps: First, we can obtain the air quality prediction result in each grid through our proposed prediction method using the ground truth data form the fixed sensors. Then the linear regression is used to calculate the air quality value of each pixel in the 1,000 pixel × 1,000 pixel image. Lastly, each pixel is assigned a color by mapping the air quality data to the pre-set color range. Our visualization highlighted the variability across different regions rather than focusing on the absolute value, which means the colors on the map represent relative values and enable us to easily and directly understand the surrounding air quality conditions.

Figure 9(b) is an example of our visualization showing the pollutants distribution of $PM_{2.5}$ at $19/12/2018\ 19:00:00$ in our divided 100 grids. The color bar at the right hand side represents the value range on map. The star, round face and triangle marks on the graph are where the fixed sensors being installed. Observing the visualization results, we find that the upper right area has higher concentration of the air pollutant factors and the center part is less polluted in general. This is because the upper area is closer to a factory area and the center region has several green parks and residential areas.



(a)

(b)

(c)

(d)

(e)

(f)

FIGURE 8. Our prediction using Random Forest Regressor (RFR), Gaussian Boosting (GB) and Support Vector Regressor (SVR) against ground truths. We show the predictions with different ranges of granularity. We find that GB performs the best among the other prediction methods.

TABLE 3. Relationship between the number of training samples in a grid vs. the prediction accuracy.

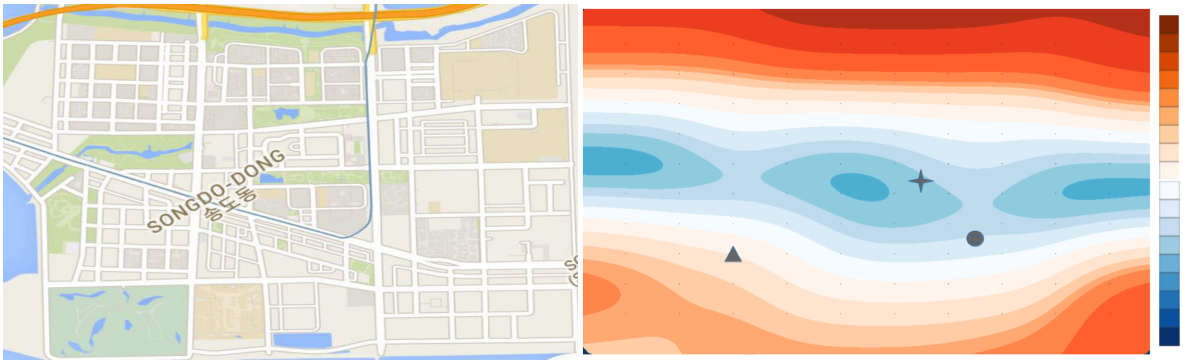
Number of samples	0-10	11-20	20-50	50-100	100-200	>200
Number of grids	76	12	5	3	1	3
RMSE	32.4	27.0	19.7	16.4	15.3	13.6

TABLE 4. Results of comparison among different interpolation methods.

	Training date	12.10-12.11	12.11-12.12	12.12-12.13	12.13-12.14
	Test date	12.12	12.13	12.14	12.15
Original baseline (no interpo.)	$RMSE(PM_{10})$	13.7	18.2	13.2	27.6
	$RMSE(PM_{2.5})$	10.3	16.8	9.5	21.6
Linear interpo. only	$RMSE(PM_{10})$	11.3	15.8	10.6	24.3
	$RMSE(PM_{2.5})$	9.0	13.9	8.2	18.7
GPR + Cauchy kernel	$RMSE(PM_{10})$	10.9	14.7	9.2	22.9
	$RMSE(PM_{2.5})$	8.3	12.3	6.8	16.6
GPR + Gaussian kernel	$RMSE(PM_{10})$	**10.4**	**14.6**	**8.7**	**22.8**
	$RMSE(PM_{2.5})$	**7.9**	**11.2**	**6.3**	**16.1**

TABLE 5. Results of comparison between our method and only using fixed sensors vs. hybrid fixed + IoT sensors.

	Training date	12.10-12.11	12.11-12.12	12.12-12.13	12.13-12.14
	Test date	12.12	12.13	12.14	12.15
$RMSE(PM_{10})$	Fixed sensors	10.9	15.4	9.1	25.6
	Hybrid sensors	**10.3**	**14.5**	**8.6**	**22.7**
$RMSE(PM_{2.5})$	Fixed sensors	8.3	12.1	6.5	17.5
	Hybrid sensors	**7.8**	**11.2**	**6.2**	**16.1**



(a) Experiment area on Google map.

(b) Heatmap of the prediction.

FIGURE 9. A visualization example of $PM_{2.5}$ at Songdo, Incheon. The air quality visualization on the right is based on the experiment area on the left. Areas in blue color means lower content of $PM_{2.5}$ and red indicates heavier pollution.

VII. DISCUSSIONS AND LIMITATIONS

It is interesting to observe that the errors are much higher in the last column in Table 4 and 5. The reason is that in our data set from Dec. 10th – Dec. 14th consists of weekdays and Dec. 15th is a Saturday, which means the air pollution patterns in the selected area are different between weekdays and weekends. The similar pattern can be also observed on Dec. 16th in Table 2, which is Sunday. By looking into the data sheet, the ground-truth data shows that in general weekends have heavier air pollution. Therefore, weekday or weekend is an important factor to consider in designing a better air pollution prediction model.

These days, deep learning techniques are widely used for classification and regression tasks. However, our initial results show that deep learning models did not perform well because of small amount of data and simple classical model performed better. For future work, after collecting more data, we plan to experiment extensively with deep learning algorithms and further incorporate different features to improve the prediction performance.

VIII. CONCLUSION

In this paper, we explored a new way to predict immediate air quality around people, by combining fixed and mobile sensors. Our experimental results show that our proposed hybrid distributed fixed and IoT sensor system is effective in predicting air quality around the people. In addition, our proposed system can be practically realizable by leveraging public transportation system such as buses as well as taxis to be equipped with IoT sensor devices to measure different areas. The predicted air quality data from our system can be served in various scenarios, such as planing for outdoor activities.

REFERENCES

[1] *Beijing'S Air Would be Step up for Smoggy Delhi*. Accessed: Jan. 26, 2014. [Online]. Available: http://https://www.nytimes.com/2014/01/26/world/asia/beijings-air-would%-be-step-up-for-smoggy-delhi.html

[2] M. Kampa and E. Castanas, "Human health effects of air pollution," *Environ. Pollut.*, vol. 151, no. 2, pp. 362–367, Jan. 2008.

[3] E. Boldo, S. Medina, A. Le Tertre, F. Hurley, H.-G. Mücke, F. Ballester, and I. Aguilera, "Apheis: Health impact assessment of long-term exposure to PM2.5 in 23 European cities," *Eur. J. Epidemiology*, vol. 21, no. 6, pp. 449–458, Jun. 2006.

[4] J. Lin, A. Zhang, W. Chen, and M. Lin, "Estimates of daily PM2.5 exposure in beijing using spatio-temporal kriging model," *Sustainability*, vol. 10, no. 8, p. 2772, 2018.

[5] Y. Jiang, L. Shang, K. Li, L. Tian, R. Piedrahita, X. Yun, O. Mansata, Q. Lv, R. P. Dick, and M. Hannigan, "MAQS: A personalized mobile sensing system for indoor air quality monitoring," in *Proc. 13th Int. Conf. Ubiquitous Comput. UbiComp*, 2011, pp. 271–280.

[6] D. Zhang and S. S. Woo, "Predicting air quality using moving sensors (poster)," in *Proc. 17th Annu. Int. Conf. Mobile Syst., Appl., Services*, Jun. 2019, pp. 604–605.

[7] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting fine-grained air quality based on big data," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining KDD*, 2015, pp. 2267–2276.

[8] M. Alvarado, F. Gonzalez, P. Erskine, D. Cliff, and D. Heuff, "A methodology to monitor airborne PM10 dust particles using a small unmanned aerial vehicle," *Sensors*, vol. 17, no. 2, p. 343, 2017.

[9] I. Kok, M. U. Simsek, and S. Ozdemir, "A deep learning model for air quality prediction in smart cities," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 1983–1990.

[10] S. Devarakonda, P. Sevusu, H. Liu, R. Liu, L. Iftode, and B. Nath, "Real-time air quality monitoring through mobile sensing in metropolitan areas," in *Proc. 2nd ACM SIGKDD Int. Workshop Urban Comput. UrbComp*, 2013, p. 15.

[11] Y.-C. Hsu, P. Dille, J. Cross, B. Dias, R. Sargent, and I. Nourbakhsh, "Community-empowered air quality monitoring system," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2017, pp. 1607–1619.

[12] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, T. Arn, J. Beutel, and L. Thiele, "Deriving high-resolution urban air pollution maps using mobile sensor nodes," *Pervas. Mobile Comput.*, vol. 16, pp. 268–285, Jan. 2015.

[13] A. C. Rai, P. Kumar, F. Pilla, A. N. Skouloudis, S. Di Sabatino, C. Ratti, A. Yasar, and D. Rickerby, "End-user perspective of low-cost sensors for outdoor air pollution monitoring," *Sci. Total Environ.*, vols. 607–608, pp. 691–705, Dec. 2017.

[14] S. R. Garzon, S. Walther, S. Pang, B. Deva, and A. Küpper, "Urban air pollution alert service for smart cities," in *Proc. 8th Int. Conf. Internet Things*, Oct. 2018, p. 8.

[15] B. Maag, Z. Zhou, and L. Thiele, "W-Air: Enabling personal air pollution monitoring on wearables," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, p. 24, 2018.

[16] R. Zhao, X. Gu, B. Xue, J. Zhang, and W. Ren, "Short period PM2.5 prediction based on multivariate linear regression model," *PLoS ONE*, vol. 13, no. 7, 2018, Art. no. e0201011.

[17] 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.

[18] X. Yi, J. Zhang, Z. Wang, T. Li, and Y. Zheng, "Deep distributed fusion network for air quality prediction," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 965–973.

[19] K. P. Moustris, P. T. Nastos, I. K. Larissi, and A. G. Paliatsos, "Application of multiple linear regression models and artificial neural networks on the surface ozone forecast in the greater athens area, greece," *Adv. Meteorol.*, vol. 2012, pp. 1–8, Jul. 2012.

[20] S. Fotouhi, M. H. Shirali-Shahreza, and A. Mohammadpour, "Concentration prediction of air pollutants in tehran," in *Proc. Int. Conf. Smart Cities Internet Things SCIOT*, 2018, pp. 1–7.

[21] C. Kim, "Place promotion and symbolic characterization of new songdo city, South Korea," *Cities*, vol. 27, no. 1, pp. 13–19, Feb. 2010.

[22] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*. Berlin, Germany: Springer, 2003, pp. 63–71.

[23] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[24] C. Cortes and V. Vapnik, "Support vector machine," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[25] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, Feb. 2002.

[26] J. Wainer, "Comparison of 14 different families of classification algorithms on 115 binary datasets," 2016, *arXiv:1606.00930*. [Online]. Available: http://arxiv.org/abs/1606.00930

[27] J. O. Ogutu, H.-P. Piepho, and T. Schulz-Streeck, "A comparison of random forests, boosting and support vector machines for genomic selection," *BMC Proc.*, vol. 5, no. S3, p. 11, c. 2011.

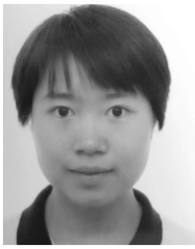
[28] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[29] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 1997, pp. 155–161.

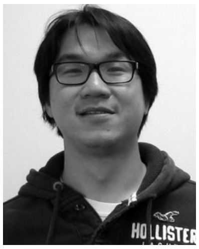
[30] N. Duffy and D. Helmbold, "Boosting methods for regression," *Mach. Learn.*, vol. 47, nos 2–3, pp. 153–200, May 2002.

[31] M. Scheuerer, R. Schaback, and M. Schlather, "Interpolation of spatial data—A stochastic or a deterministic problem?" *Eur. J. Appl. Math.*, vol. 24, no. 4, pp. 601–629, Aug. 2013.

[32] P. Völgyesi, A. Nádas, X. Koutsoukos, Á. Lédeczi, "Air quality monitoring with SensorMap," in *Proc. Int. Conf. Inf. Process. Sensor Netw. (ipsn)*, Apr. 2008, pp. 529–530.



DAN ZHANG received the B.E. degree in computer science and technology from the University of Electronic Science and Technology of China, China, and the M.S. degree from the Department of Computer Science, State University of Korea, South Korea. She is currently pursuing the Ph.D. degree with the Department of Computer Science, Stony Brook University. Her current research interests include data analysis, machine learning, deep learning, and visualization.



SIMON S. WOO received the B.S. degree in electrical engineering from the University of Washington (UW), Seattle, the M.S. degree in electrical and computer engineering from the University of California at San Diego (UCSD), and the M.S. and Ph.D. degrees in computer science from the University of Southern California (USC), Los Angeles. He was a member of technical staff (technologist), for nine years, at the NASA's Jet Propulsion Lab (JPL), Pasadena, CA, USA, conducting research in satellite communications, networking, and cybersecurity areas. He also worked at Intel Corporation and Verisign Research Lab. Since 2017, he was a tenure-track Assistant Professor at SUNY Korea, South Korea, and a Research Assistant Professor at Stony Brook University. He is currently a tenure-track Assistant Professor at the SKKU Institute for Convergence and the Department of Applied Data Science, Sungkyunkwan University, Suwon, South Korea.