A Comparative Study on Air Quality Analysis Through DNN by SVM, K - Means and Naive Bayes Algorithms

R.Amulraju[1] D. Ashok Kumar[2] T.Vithyaa[3]
Department of Computer Science, Government Arts College, Kulithalai-639120. Karur District, Tamilnadu

Abstract — *Air Quality may be a major concern round the world. It's full of a large vary of natural and human influences. The foremost necessary of the natural influences area unit geologic, hydrological and environmental condition, since these have an effect on the standard of Air. To invoke a deep neural network (DNN)-based approach (entitled Deep Air), that consists of a spatial transformation part and a deep distributed fusion network. Considering air pollutants' spatial correlations, the previous part converts the spatial thin air quality knowledge into a homogenous input to simulate the waste product sources. The latter network adopts a neural distributed design to fuse heterogeneous urban knowledge for at the same time capturing the factors touching air quality, e.g. environmental condition. To Deployed Deep Air in our pollution Prediction system, providing fine-grained air quality forecast. Additionally we tend to confirm the precise results and analysis of the contaminated contents victimization K-Means cluster, SVM Classifier and Naive Bayes. Comparison the contaminated content results with these three processes the K-Means provides the proper result and determines the precise output with facilitate of the datasets. The collected datasets area unit pre-processed and classified to induce the proper results. Finally the results area unit manipulated in associate graph format, that exposes the ends up in associate correct manner.*

Key terms: *Air Quality, K-means, SVM, Naïve Bayes, Precision, Recall, F-Measure*

I. INTRODUCTION

The interpolation, prediction, and have analysis of fine-gained air quality area unit 3 vital topics within the space of urban air computing. a decent interpolation solves the matter that there are a unit restricted air-quality-monitor-stations whose distribution is uneven in a very city; a certain prediction provides valuable info to shield humans from being broken by air pollution; an inexpensive feature analysis reveals the most relevant factors to the variation of air quality. In general, the solutions to those topics will extract very helpful info to support pollution management, and consequently generate nice social group and technical impacts.
However, there exist many challenges for urban air computing because the connected information has some special characteristics. First, since there are a unit scant air quality-monitor stations in a very town because of the high value of building and maintaining such a station, it's dearly won to get labeled coaching samples once coping with fine gained air quality. Second, the labeled information of the air quality-monitor-stations area unit incomplete, and there exist numerous missing labels of the historical information in it slow periods for a few stations. The explanation for the unfinished labels is expounded to the air quality monitor devices. In general, every station solely has one monitor device that must be maintained at intervals, so there'll be no outputs for the station once the device is being maintained, recalibrated, or has different issues. Third, the styles of urban air connected information area unit numerous for the event of knowledge acquisition technologies. However, there's not a universally accepted judgment to reveal the most causes of the prevalence and dissipation of pollution, particularly the pollution of PM2:5. Hence, it's laborious to grasp that what styles of information area unit the most relevant options for interpolation and prediction, and therefore the key factors for surroundings departments to stop and management pollution.

Data science is that the study of wherever data comes from, what it represents and the way it may be was a valuable resource within the creation of business and IT ways. Mining giant amounts of structured and unstructured information to spot patterns will facilitate a company rein in prices, increase efficiencies, acknowledge new market opportunities and increase the organization's competitive advantage. The information science field employs arithmetic, statistics and computing disciplines, and incorporates techniques like machine learning, cluster analysis, data processing and visual image.

II. RELATEDWORKS

Ping-Wei Soh [5] Air pollution has become a very significant issue, with stuff having a considerably bigger impact on human health than alternative contaminants. The little diameter of fine stuff (PM2.5) permits it to penetrate

Deep into the alveoli as way because the bronchioles, officious with a gas exchange among the lungs. Long exposure to stuff has been shown to cause the upset, respiratory illness, and increase the chance of respiratory organ cancers. Therefore, statement air quality has conjointly become vital to assist guide individual actions. This paper aims to forecast air quality for up to forty eight h employing a combination of multiple neural networks, together with a synthetic neural network, a convolution neural network, and a long-short-term memory to extract spatial-temporal relations. The planned prognosticative model considers varied meteorology knowledge from the last few hours further as info associated with the elevation area to extract piece of ground impact on air quality. The model includes trends from multiple locations, extracted from correlations between adjacent locations, and among similar locations within the temporal domain. Experiments using Taiwan and Peiping knowledge sets show that the planned model achieves glorious performance and outperforms current progressive ways.

Susan A Perlin [1] we recognized that a lot of health outcomes are related to pollution, however during this project launched by the North American country Environmental Protection Agency, the intent was to assess the role of exposure to close air pollutants as risk factors just for metabolic process effects in youngsters. The NHANES-III information may be a valuable resource for assessing children's metabolic process health and bound risk factors however lacks observance information to estimate subjects' exposures to close air pollutants. Since the Seventies, Environmental Protection Agency has frequently monitored levels of many close air pollutants across the country and this information is also wont to estimate NHANES subject's exposure to close air pollutants. The primary stage of the project eventually evolved into assessing totally different estimation strategies before adopting the estimates to gauge metabolic process health. Specifically, this paper describes an attempt exploitation EPA's AIRS observance information to estimate gas and PM10 levels at census block teams. we have a tendency to restricted those block teams to counties visited by NHANES-III to create the project a lot of manageable and apply four totally different interpolation strategies to the observance information to derive air concentration levels. Then we have a tendency to examine method-specific variations in concentration levels and confirm conditions below that totally strategies turn out considerably different concentration values. we discover that different interpolation strategies don't turn out dramatically different estimations in most elements of the North American country wherever monitor density was comparatively low. However, in areas wherever monitor density was comparatively high (i.e., California), we discover substantial variations in exposure estimates across the interpolation strategies. Our results provide some insights into terms of exploitation the Environmental Protection Agency observance information for the chosen abstraction interpolation strategies.

Gehrig, Y. [8]Using one year of aerosol optical thickness

(AOT) retrievals from the MODerate resolution Imaging Spectro-radiometer (MODIS) on board NASA's Terra and cobalt blue satellite alongside ground measurements of PM2.5 mass concentration, we tend to assess stuff air quality over completely different locations across the worldwide urban areas cover twenty six locations in Sydney, Delhi, Hong Kong, the big apple town and Suisse. Associate degree empirical relationship between AOT and PM2.5 mass is obtained and results show that there's a wonderful correlation between the bin-averaged daily mean satellite and ground-based values with a linear coefficient of correlation of zero.96. Mistreatment meteorological and alternative subsidiary datasets, we tend to assess the results of wind speed, cloudiness, and mix height (MH) on stuff (PM) air quality and conclude that these information area unit necessary to more apply satellite information for air quality analysis. Our study clearly demonstrates that satellite-derived AOT could be a sensible surrogate for watching PM air quality over the planet. However, our analysis shows that the PM2.5–AOT relationship powerfully depends on aerosol concentrations, close ratio (RH), uncompleted cloudiness and height of the blending layer. Highest correlation between MODIS AOT and PM2.5 mass is found underneath clear sky conditions with but 40–50% RH and once part MH ranges from a hundred to two hundred m. Future remote sensing sensors like Cloud-Aerosol measuring instrument and Infrared scout Satellite Observations (CALIPSO) that have the potential to supply vertical distribution of aerosols can more enhance our ability to observe and forecast pollution. This study is among the primary to look at the link between satellite and ground measurements over many international locations.

Ultra-fine particles with aerodynamic diameter smaller than 2.5 microns, namely Particulate Matter 2.5 (PM 2.5), are capable of penetrating the lung cells and circulating the circulatory system, and compose a major health threat to people. Although the government is publishing the outdoor PM2.5 concentration on an hourly basis, the indoor PM 2.5 concentration, to which

most people expose for most of their everyday life time, remains unsupervised. The high cost of the professional PM 2.5 measuring equipments, which utilize filtering and direct mass measuring methodology, prevents the indoor air quality to be monitored pervasively. We designed and implemented PiMi air box, a cost-effective portable sensor, which is able to estimate the PM 2.5 mass concentration with satisfactory accuracy. The PiMi air boxes adopt the low-cost optical particle counting technology and convert the particle counts into PM 2.5 mass concentrations via empirical diameter-distribution and density of particulate matters. The errors introduced by the individuality of the low-cost particle counters are offset during a machine-learning-based calibration procedure for each single unit. The PiMi air box enjoys a stunning cost reduction by a factor of 1,000 comparing to professional equipments, and still maintains an satisfactory level of accuracy for everyday life air quality measurement. Together with embedded Bluetooth connectivity and Smartphone APPs, PiMi air box is well-suited for massive crowd-sourced indoor air-quality
Monitoring research.

Z. Shan, [13] In this paper, we have a tendency to forecast the reading of Associate in Nursing air quality watching station within the next forty eight hours, employing a information-driven methodology that considers the present earth science data, weather forecasts, and therefore the air quality information of the station which of different stations inside many hundred kilometres to the station. Our prognostic model is comprised of 4 major components: (1) a linear regression-based temporal predictor to model the native issue of air quality, (2) a neural network based spatial predictor modelling the worldwide factors, (3) a dynamic individual combining the predictions of the spatial and temporal predictors in keeping with the earth science information, Associate in Nursing (4) an inflection predictor to capture the explosive changes of air quality. We have a tendency to evaluate our model with the information of forty three cities in China, surpassing the results of multiple baseline ways. We've deployed a system in Chinese Ministry of Environmental Protection, providing 48-hour fine-grained air quality forecasts for four major Chinese cities each hour. The forecast operate is additionally enabled on Microsoft Bing Map and MS cloud platform Azure. Our technology is general and might be applied globally for different cities.

Y. Zheng [12] Traditional data processing sometimes deals with information from one domain. Within the huge information era, we have a tendency to face a diversity of datasets from totally different sources in several domains. These datasets carries with it multiple modalities, every of that contains a totally different illustration, distribution, scale, and density. The way to unlock the ability of knowledge from multiple disparate (but probably connected) datasets is preponderating in huge data analysis, basically identifying huge information from ancient data processing tasks. This entails advanced techniques which will fuse data from varied datasets organically in an exceedingly machine learning and data processing task. This paper summarizes the info fusion methodologies, classifying them into 3 categories: stage-based, feature level-based, and linguistics meaning-based information fusion strategies. The last class of information fusion strategies is any divided into four groups: multi-view learning-based, similarity-based, and probabilistic dependency-based, and transfer learning-based strategies. These strategies specialize in data fusion instead of schema mapping and information merging, considerably identifying between cross-domain information fusion and ancient information fusion studied within the info community. This paper doesn't solely introduce high-level principles of every class of strategies; however conjointly offer examples during which these techniques square measure accustomed handle real huge information issues. Additionally, this paper positions existing works in an exceedingly framework, exploring the connection and distinction between totally different information fusion strategies. This paper can facilitate a good vary of communities realize an answer for information fusion in huge information comes.

D. Basak, [11] Support Vector Regression (SVR), a class for Support Vector Machine (SVM) makes an attempt to reduce the generalization error certain therefore on deliver the goods generalized performance. Regression is that of finding a operate that approximates mapping from Associate in Nursing input domain to the important numbers on the premise of a coaching sample. Support vector regression is that the natural extension of huge margin kernel strategies used for classification to multivariate analysis. On account of steady increase in paper demand, the forecast on demand and provide of pulp wood is taken into account to boost the socio economic development of Asian country.

III. PROBLEM & MODEL DESCRIPTION

The main goal of this system is to predict Air quality using data mining technique such as Naive Bayesian Algorithm. Raw data set is used and then pre-processed and transformed the data set. To apply clustering algorithm such as K means algorithm and then apply the data mining classification technique such as Naïve Bayes algorithm and SVM on the transformed data set. After applying the data mining algorithm, Air quality is predicted and then accuracy is calculated.

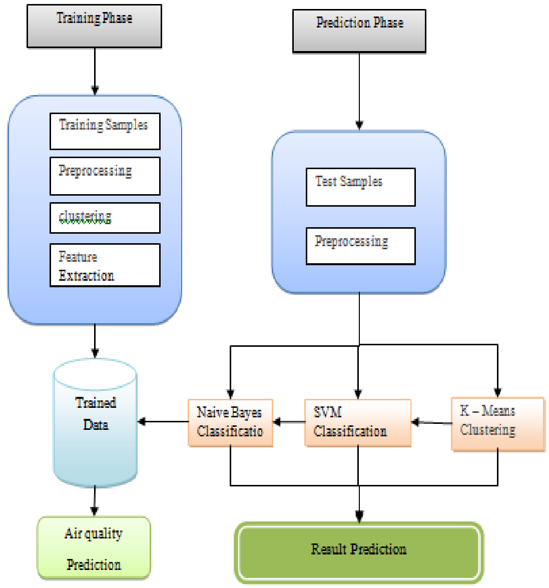

Fig 1.DNN Architecture

IV. SYSTEM IMPLEMENTATION

A. Dataset Acquisition

In this module, upload the datasets. Gather the data from data centers. The composed data is pre-processed and stored in the knowledge base to build the model.

B. Pre-processing

Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine projects. Data-gathering methods are often loosely guarded, resulting in out-of-range values, unfeasible data combinations, missing values, etc. Analyzing data that has not been suspiciously screened for such problems can produce misleading results.

C. Clustering

Clustering is a technique in data mining to find interesting patterns in a given dataset .The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters information's into k groups, where k is considered as an input factor. It then assigns each information's to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then more computed and the process begins again. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical data and related fields. K-Means algorithm is a divisive, unordered method of defining clusters.

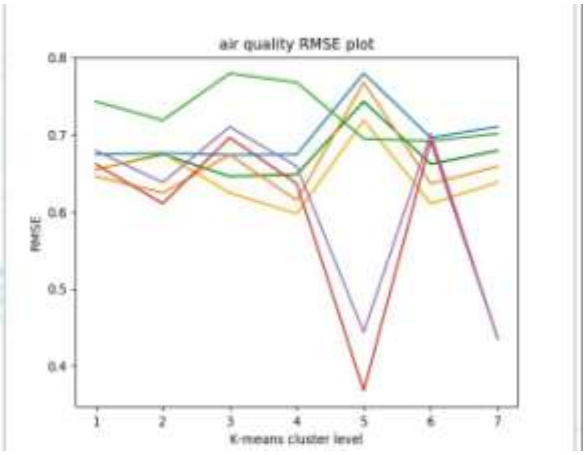

Fig 2: Air Quality RMSE Plot

D. Feature Selection

In this module is used to select the features of the given dataset. Attribute selection was performed to determine the subset of features that were highly correlated with the class while having low inter correlation.

Classification

Naïve Bayesian

The Naïve Bayesian Classification Algorithm represents a statistical method with supervised learning method for classification. Assumes a probabilistic model which allows us to resolve the diagnostic and predictive problems. Bayes classification has been proposed which is based on Bayes rule of conditional probability. Naïve Bayesian rule is a technique used to estimate the likelihood of a property from the given data set. The approach is called "naïve" for the reason that assumes the independence between the various attribute values. Bayesian classification can be seen as both a descriptive and a predictive type of algorithm. The probabilities are expressive and used to predict the class membership for a target duple.

SVM

Support Vector Machine (SVM) is a machine learning tool that is based on the idea of large boundary data classification. The tool has strong speculative foundation and the classification algorithms based on it give good simplification performance. Standard implementations, though provide fine classification accuracy, are slow and do not scale well. Hence they cannot be applied to large-scale data mining applications. They typically need large number of support vectors. Hence the training as well as the classification times is high.

- Input: Input data matrix, class information

- Output: Set of Basis vectors Begin Repeat

For every candidate example - examples not in current set of BVs Include it in the model efficiently Observe the generalization performance on the remaining points end for candidate examples add that point to the BVs list that gave better test error Till the stopping criterion End.

Precision

In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the query:

For example, for a text search on a set of documents, precision is the number of correct results divided by the number of all returned results.

Precision takes all retrieved documents into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision at n or P@n.

Precision is used with recall, the percent of all relevant documents that is returned by the search. The two measures are sometimes used together in the F1 Score (or f-measure) to provide a single measurement for a system.

Clustering algorithm for prediction of air quality. Thus Air
Quality prediction system successfully diagnoses the data
and predicts the Air quality. The results thus obtained shows
That K- Means clustering algorithm provides 86.58% of accuracy with minimum time.

Dataset Description

Our proposed scheme and datasets are related to the air and pollution causing gases and molecules that are presented in the environment. The dataset contains the labeled data for 9358 iterations that are related to the air

Note that the meaning and usage of "precision" in the field of information retrieval differs from the definition of accuracy and precision within other branches of science and technology.

Recall

In information retrieval, recall is the fraction of the relevant documents that are successfully retrieved.

For example, for a text search on a set of documents, recall is the number of correct results divided by the number of results that should have been returned.

In binary classification, recall is called sensitivity. It can be viewed as the probability that a relevant document is retrieved by the query.

It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough but one needs to measure the number of non-relevant documents also, for example by also computing the
Precision.

V. RESULT EXPECTED AND OBTAINED

In the comparative work naviebayes, SVM algorithm is used to classify and the data set, and it also used K-means

pollution dataset, which we divide into training sets that assume the values from 0 – 9357 to gain the mean and accuracy level, at the same time the SVM and Naïve Bayes concludes only part values in the data set and misses some values which results in inaccuracy. Note that the dataset does not have any labeled nodes. For each value it consists of numeric values. The accuracy level values are provided in the table by considering the K – Means clustering which distributes the accurate of the polluted content present in the environment and classifies with the other prediction algorithms or reduce the imbalance in the system that has been proposed.

1. Dataset Table

		CO (GT)	**PT(CO)**	**RF**	**IF**
Testing phase	**Count**	9357	9357	9357	9357
	Mean	-34.20	1048	39.48	-6.83
	Standard	77.65	329	51.21	38.97
	Min	-200	-200	-200	-200
	25%	0.600	921	34.05	0.69
	50%	1.50	1052	48.55	0.97
	75%	2.60	1221	61.87	1.29

	Max	11.90	2039	88.72	2.23
Training phase		**CO (GT)**	**PT(CO)**	**RF**	**IF**
	Count	7674	8991	8991	8991
	Mean	2.15	1099	49.23	1.02
	Standard	1.45	217	17.31	0.40
	Min	0.10	647	9.17	0.18
	25%	1.10	936	35.81	0.73
	50%	1.80	1063	49.55	0.99
	75%	2.90	1231	62.50	1.31
	Max	11.90	2039	88.72	2.23

Table 1.Dataset Table.

RF **--** Temperature, wind strength, NO2, CO2 etc,   IF **--**Barometric Pressure, Humidity, O3
PT – PrecipitationTemperature   CO **--** Carbon di Oxide

Accuracy Level Prediction Proposed Method

Precision	0.7516	**T P**	115	75.16%
	0.9358	**TN**	5	93.58%
Recall	0.9583	**FP**	38	95.83%
	0.6576	**F N**	73	65.76%
Accuracy				81.38%

TP - True Positive   **FP - False Positive**
TN - True Negative   **FN - False Negative**

Precision Recalland F-Measure

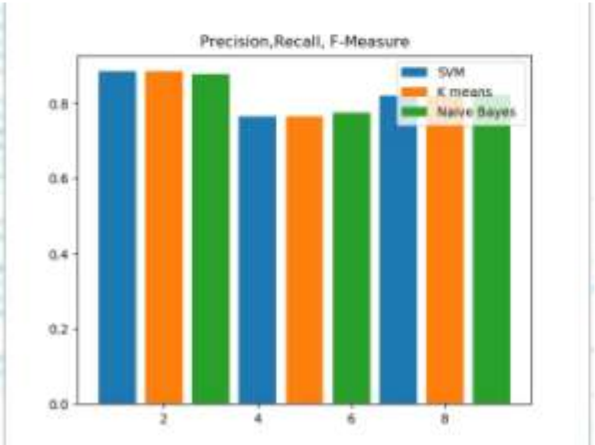

Fig 3: Precision, Recall,F-Measure.

Precision Recalland F-Measure are then defined as

$$\boldsymbol{Preceision = TP/(TP + FP)}$$

$$\boldsymbol{Recall = TP/(TP + FN)}$$

F = 2*[(Precision*Recall) / (Precision +Recall)]

Accuracy

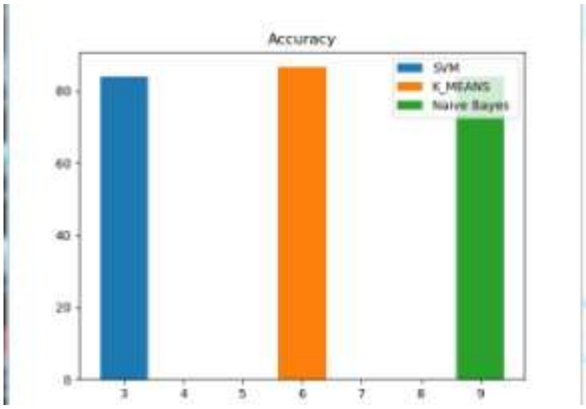

Fig 4:Accurcy Graph

$$\mathbf{Accuracy = TP + TN/(TP + TN + FP + FN)}$$

2. Compression Table

Methods	P	R	F	ACCURACY
Naive Bayes	0.87	0.774	0.8229	83.98%
SVM	0.88	0.765	0.82	83.98%,
K-Means	0.88	0.765	0.8212	86.58%,

Table 2.Comparission Table

Accuracy Rate

The below figure shows the Accuracy rate having value 86 % which is higher than the SVM (Support Vector Machine) Algorithm and Naïve Bayesian Algorithm.

Whereas,

P – Precision
R- Recall
F- F-Measure
A- Accuracy

In this below table is represent the comparison between the three classifier algorithms such as Support Vector Machine , Naive Bayes and K-Means. In the our method k-Means Clustering provide the accuracy level (86.58%) is higher than the Support Vector Machine and Naïve Bayes algorithms.

VI. CONCLUSION

Data Science applications are used enormously in the industrial field to detect Air quality based on the data set and the attributes provided. Researchers have been investigating applying different data mining techniques to help research professionals in the prediction of air quality. In the comparative work naviebayes, SVM algorithm is used to classify and the data set, and it also used K-means clustering algorithm for prediction of air quality. Thus Air quality prediction system successfully diagnoses the data and predicts the Air quality. The results thus obtained shows that K- Means clustering algorithm provides 86.58% of accuracy with minimum time.

Reference

[1] D. W. Wong, L. Yuan, and S. A. Perlin, "Comparison ofspatialinterpo-lation methods for the estimation of air quality data,"Journalof Exposure Science and Environmental Epidemiology, vol. 14, no. 5,, 2004.

[2] R. G. Baraniuk, "Compressive sensing, sensing," IEEE Signal Processing Magazine Magazine, vol. 24, no. 4, 2007.

[3] S. S. Haykin, S. S. Haykin, S. S. Haykin, and S. S. Haykin, Neural networks and learning machines. Pearson Education Upper Saddle River, 2009,

[4] J. Schwartz, "Lung function and chronic exposure to air pollution: A crosscross-sectional analysis of NHANES II, II," Environmental Research Research, vol. 50,no. 2, pp. 309 –321, 1989.

[5] Ping-Wei Soh, Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations, , vol. 50,no. 2, pp. 309 –321, 1989.

[6] L. G. Chestnut, J. Schwartz, D. A. Savitz, and C. M. Burchfiel, "Pulmo-nary function and ambient particulate matter: epidemiological evidence from NHANES I, I," Archives of Environmental Health: An International Journal , vol. 46, no. 3, 1991.

[7] D. K. Jha, M. Sabesan , A. Das, N. Vinithkumar , and R. "Evaluation of interpolation technique for air quality parameters in port blairblair, india ," Universal Journal of Environmental Research and TechnologyTechnology, vol. 1, no. 3, 2011.

[8] P. Gupta, S. A. Christopher, J. Wang, R. Gehrig, Y. Lee, and N. Kumar, "Satellite remote sensing of particulate matter and air quality assessment over global cities, cities," Atmospheric Environment , vol. 40, no. 30, 2006.

[9] X. Yu, Y. Liu, Y. Zhu, W. Feng, L. Zhang, H. F. Rashvand, and V. O. K. Li, "Efficient sampling and compressive sensing for urban monitoring vehicular sensor networks, networks," IET Wireless Sensor Systems , vol. 2, 2012.

[10] L. Li, Y. Zheng, and L. Zhang, "Demonstration abstract: Pimi air box: a costcost-effective sensor for par ticipatory indoor quality monitoring, monitoring," in Pro- ceedings of the 13th IEEE International Symposium on Information Processing in Sensor Networks , 2014.

[11] D. Basak, S. Pal, and D. C. Patranabis. "Support vector regression." Neu-ral Information Processing-Letters and Reviews 11, no. 10 2007.

[12]Y. Zheng. "Methodologies for cross-domain data fusion:An overview." IEEE transactions on big data 1, no. 1 2015.

[13] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li. "Forecasting Fine-Grained Air Quality Based on Big Data." 2015.

[14] N. Cressie and C. K. Wikle, Statistics for Spatio-Temporal Data. Wiley, 2011

[15] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, vol. 23, no. 19, pp.2507–2517, 2007.