# AIR QUALITY ANALYSIS USING MACHINE LEARNING

[1]Bitty Cleatus, [2]Heeba Mouinuddin, [3]Khushi Singh, [4]Lakshmi Madhumitha P, [5]Ms.Indu
[1]Student, [2]Student, [3] Student, [4] Student, [5]Asst Professor
[1]Information Science and Engineering
[1]The Oxford College of Engineering, Bangalore,
India

***Abstract:*** With the rapid development of urbanization and industrialization, many developing countries are suffering from heavy air pollution. Current air quality prediction methods mainly use shallow models; however, these methods produce unsatisfactory results. The three important topics involved in urban air computing are interpolation, prediction, and feature analysis of fine-gained air quality and the solutions to these topics can provide extremely useful information to support air pollution control, and consequently generate great societal and technical impacts. The related existing work solves the three problems separately by different models. The proposed approach utilizes the unlabelled spatio-temporal data to improve the performance of the interpolation and the prediction, and performs feature selection and association analysis, results reveals the variation in the air quality. This paper focuses on applicability of machine learning algorithms in operational conditions of air quality monitoring for predicting the daily peak concentration of a major photochemical pollutant from point measurements of a local monitoring station for the smaller places of the cities. The aim of the research reported here is the investigation of applicability of machine learning techniques for air quality forecasting in operational conditions.

***Key Words:*** **Interpolation, Prediction, Feature analysis, Air quality, Semi supervised learning, unlabelled spatio-temporal data**.

## I.INTRODUCTION

The interpolation, prediction, and feature analysis of fine-gained air quality are three important verticals of the area of urban air computing. Interpolation targets to solve the problem that there are limited air-quality-monitor-stations. A precise prediction provides valuable insight to protect humans from ill effect of the air pollution. A reasonable feature analysis provides details of the variation in the air quality.

There are several challenges for urban air computing as the related data have some special characteristics.

There are insufficient air-quality-monitor stations in a city due to the high cost of building and maintaining. Also, it is expensive to obtain labelled training samples for the fine-gained air quality.

There are lot of missing and incomplete data from the air-quality-monitor-stations. This could be because of the periodical maintenance, frequency of collection and other issues.

It's difficult to identify the kind of data that are relevant features for interpolation and prediction, and the key factors for environment departments to prevent and control air pollution. This is because there is not clearly accepted factor for the cause of air pollution. This paper addresses all these challenges by utilizing the information contained in the unlabelled data and the spatio-temporal data, and performing feature selection and association analysis for the urban air related data.

## II.PROBLEM STATEMENT

The increase in the pollution, a smaller number of devices monitoring the air quality but forward various health hazards and poorly managed air pollution. The interpolation, prediction, and feature analysis of fine-gained air quality are three important verticals in the area of air pollution analysis. The solutions to these three verticals can provide extremely useful insight to support air pollution control, and consequently generate great societal and technical impacts. Most of the existing work solves the three problems separately by different models.

## III.PROPOSED SYSTEM

The paper is motivated to address all these challenges by utilizing the information contained in the unlabelled data and the spatio-temporal data, and performing feature selection and association analysis for the urban air related data. Though labelled data are difficult or expensive to obtain, large amounts of unlabelled examples can often be gathered cheaply. In general, unlabelled data can help in providing information to better exploit the geometric structure of the data. Moreover, most of the urban air related data contain both space and time information.

The first step is determining a subset of the initial features is called feature selection. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

The feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps.

When the input data to an algorithm is too large to be processed and it is suspected to be redundant, then it can be transformed into a reduced set of features.

The learning algorithm finds patterns in the training data that map the input data attributes to the target, and it outputs an ML model that captures these patterns. Once the build model is tested then we will pass real time data for the prediction. Once prediction is done then we will analyse the output to find out the crucial information.

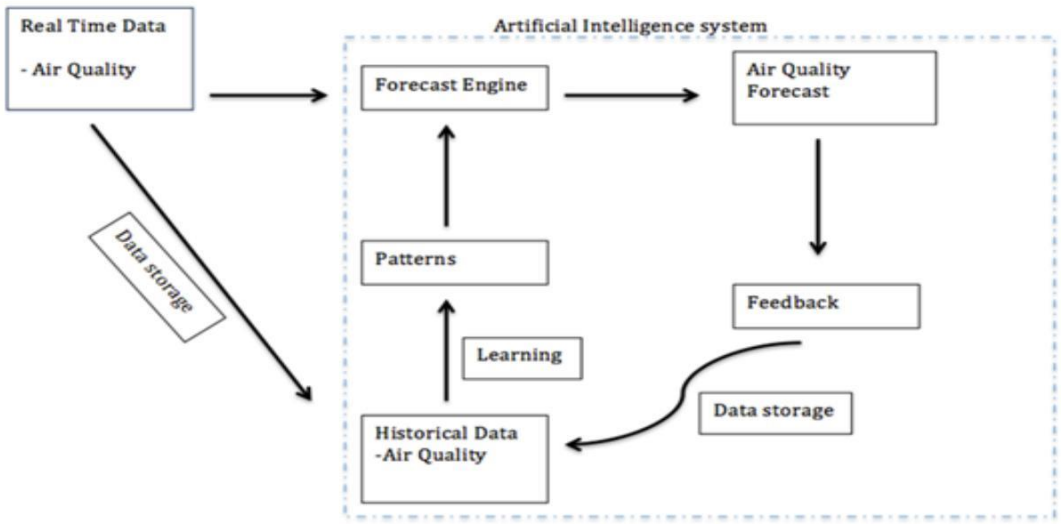


*Fig: Design Consideration of the Proposed System*

## III.RESEARCH METHODOLOGY

### 3.1 Modules Description

**Pre-processing of Captured Image**

Real Time air quality data collection

The primary data collected from the online sources remains in the raw form of statements, digits and qualitative terms. The raw data contains error, omissions and inconsistencies. Various techniques have been incorporated for the pre-processing & cleaning of the raw data. Post the pre-processing of the raw data it is feed to forecast engine and persisted in Historical Data.

Forecast Engine

The forecast engine does the feature selection. The features extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps.

Air Quality Forecast

Based on the real time air quality data and model trained, this layer does the prediction. Once prediction is done then we will analyze the output to find out the crucial information.

Feedback

The prediction is further feedback and persisted to the historical data. This is later used by the Patterns module for the analysis and training of the Machine Learning model for the prediction.

Patterns

This module involves the training the model based on the historical data and feedback from the forecast engine. The learning algorithm finds patterns in the training data that map the input data attributes to the target, and it outputs an ML model that captures these patterns. The build model is further tested for the accuracy.

### 3.2 Result Analysis

In this section, the performance of the system is checked with respect to the real time air quality statistics collected.


*Fig: Accuracy of a sample air quality data*

The above pie chart depicts the accuracy of the proposed system in predicting the urban air quality.

### 3.3 Future Work

The current algorithm uses the Random forest algorithm for classification of the air quality. In future work, the artificial neural network can be implemented for the classification problem.

## IV. Conclusion

In this project the three important topics in the area of urban air computing: the interpolation, prediction, and feature analysis of fine-gained air quality has been computed. The solutions to these topics can provide crucial information to support air pollution control, and consequently generate great societal and technical impacts. Most existing efforts focus on solving the three problems separately by establishing different models. In this paper, we develop a general and effective approach to unify the interpolation, prediction, feature selection and analysis of the fine-grained air quality into one model. In order to improve the performance of interpolation and prediction, we utilize the intrinsic characteristics of the spatio-temporal data and the information contained in the unlabeled data by embedding spatio-temporal semisupervised learning on the output layer of neural network.

## V. Acknowledgment

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of people who made it possible whose constant guidance and encouragement crowned our effort with success. We would like to express our gratitude to Dr. Praveena Gowda, Principal, The Oxford College of Engineering for providing us a congenial environment and surrounding to work in. Our hearty thanks to Dr. R. Kanagavalli, Professor & Head, Department of Information Science and Engineering, The Oxford College of Engineering for her encouragement and support. Guidance and deadlines play a very important role in successful completion of the project report on time. We convey our gratitude to Ms. Indu K S, Assistant Professor, Department of Information Science and Engineering for having constantly monitored the completion of the Project Report and setting up precise deadlines. Finally a note of thanks to the Department of Information Science and Engineering, The Oxford College of Engineering, both teaching and non-teaching staff for their cooperation extended to us.

## VI. REFERENCES

[1] Zhongang Qi, Tianchun Wang, Guojie Song, Weisong Hu, Xi Li∗, Zhongfei (Mark) Zhang "Deep Air Learning: Interpolation, Prediction, and Feature Analysis of Fine-grained Air Quality", IEEE transactions on knowledge and data engineering 2018.

[2] Hsun-Ping Hsieh, Shou-De Lin, Yu Zheng "Inferring Air Quality for Station Location Recommendation Based on Urban Big Data", in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '15, 2015, pp. 437– 446.

[3] Dora Erdos, Vatche Ishakian, Andrei Lapets, Evimaria Terzi, Azer Bestavros "The Filter Placement Problem and its Application to Minimizing Information Multiplicity ", vol. 2, no. 1, pp. 1–127, 2009.

[4] Lixin Li, Xingyou Zhang, James B.Holt, Jie Tian, Reinhard Piltner "Spatiotemporal Interpolation Methods for Air Pollution Exposure", in Symposium on Abstraction, Reformulation, and Approximation, 2011.

[5] David Hasenfratz, Olga Saukh, Silvan Sturzenegger, Lothar Thiele "Participatory Air Pollution Monitoring Using Smartphones".

[6] Dmytro Karamshuk, Anastasios Noulas, Salvatore Scellatu, Vincenzo Nicosia, Cecilia Mascolo "Geo-Spotting: Mining Online Location-based Services for Optimal Retail Store Placement"

[7] Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, Tianrui Li "Forecasting fine grained air-quality based on big data", in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '15, 2015.

[8] Salimol Thomas & Robert B. Jacko "Model for Forecasting Expressway Fine Particulate Matter and Carbon Monoxide Concentration: Application of Regression and Neural Network Models"

[9] Yu Zheng, Furui Liu, Hsun-Ping Hsieh "U-Air- When Urban Air Quality Inference Meets Big Data", in Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '13, 2013, pp. 1436–1444.