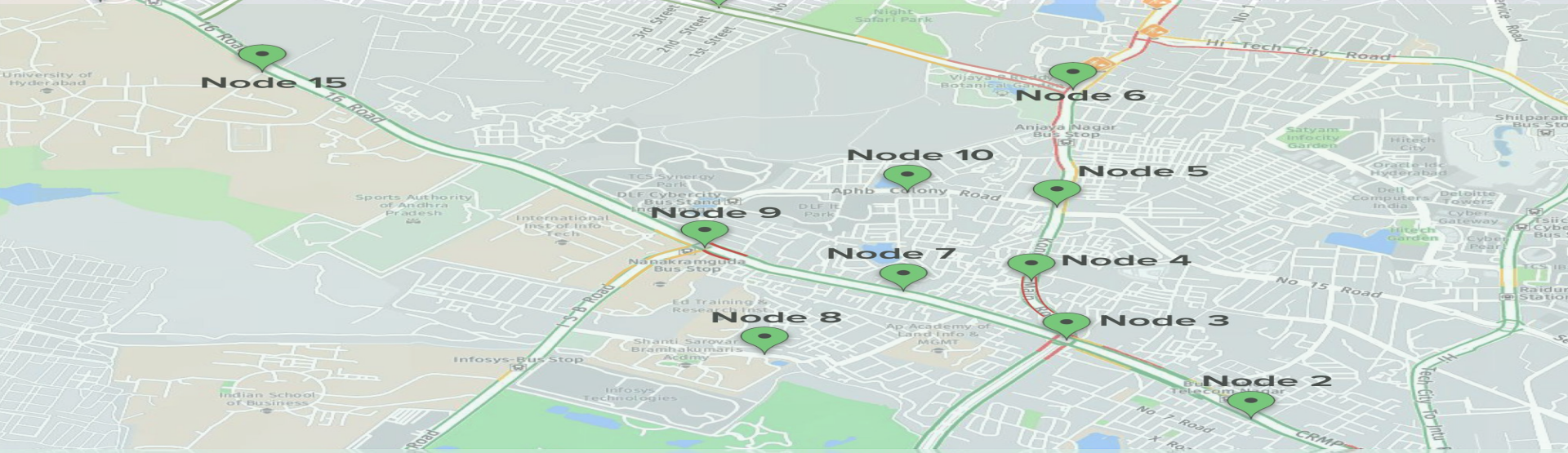


IoT and ML-based AQI Estimation using Real-time Traffic Data

Nitin Nilesh, Ritik Yelekar, Ayu Parmar, Sachin Chaudhari

Signal Processing and Communication Research Center,
IIIT Hyderabad, India

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Motivation 

- Air quality data from scientific grade devices are accurate but has the limitation of scalability as these devices are costly, bulky, and difficult to maintain.

- On the other hand, low-cost sensors have low accuracy, need calibration seasonally, and have limited lifetime (few months or years).

- For these reasons, it is desirable to have a method which does not depend on sensors.

- Therefore, we propose a method using real-time traffic data and weather parameters to determine the AQI.

Contribution 

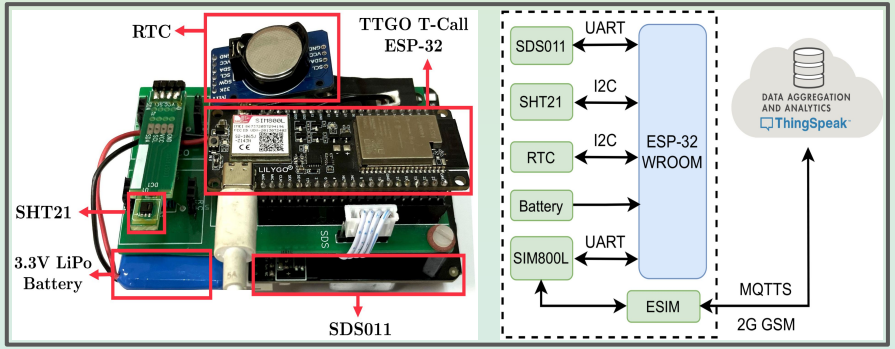
- An **IoT and ML-based methodology** is proposed to **estimate the real-time AQI** into five levels **using real-time traffic data and weather parameters**.

- A completely new **rich traffic dataset has been collected** containing approximately **210,000 data points**, including traffic information (such as the **mobility rate of the traffic**), weather information (**temperature and relative humidity**) and co-located ground truth PM values. The dataset contains samples across the **15 different locations in Hyderabad** from Jan'22 - May'22.

- A simple yet effective **ML algorithm** is used to **estimate the AQI level**, which enables the whole pipeline to be **fast and real-time** with minimal processing.

- The proposed method achieved an overall **accuracy of 82.60%** with an **F1-Score of 83.67%**. We also show the results on individual traffic locations to better understand the scenario.

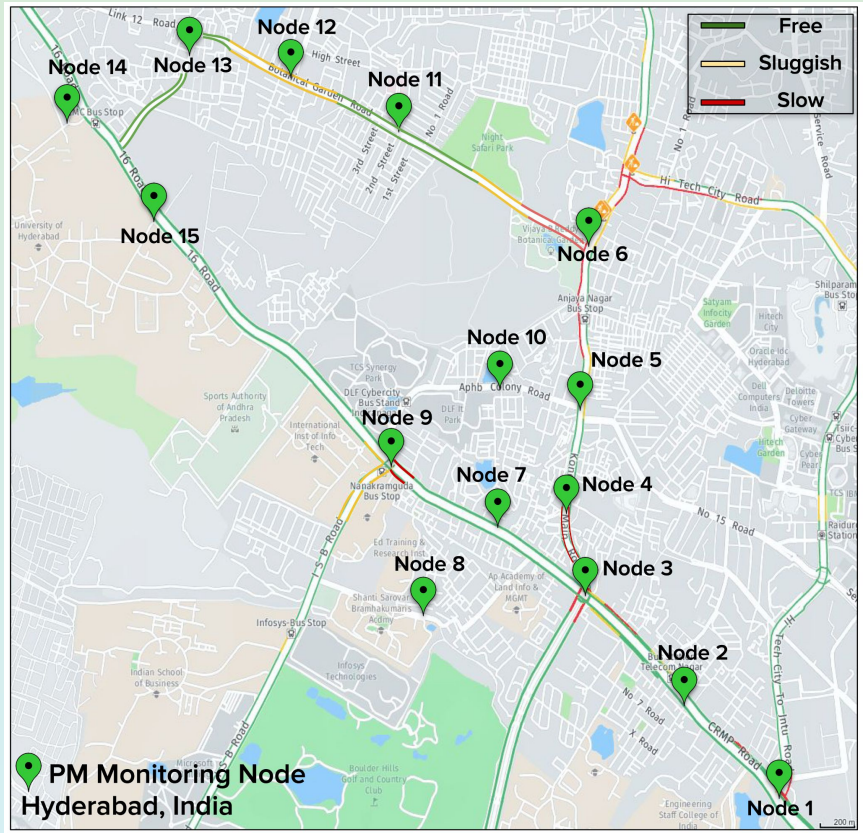
IoT Network Setup



Block architecture and the circuit board of the IoT PM monitoring node deployed in the main road and junctions.

- TTGO T-Call ESP32 based microcontroller.
- SDS011 sensor for $PM_{2.5}$ and PM_{10} measurement.
- SHT21 sensor for temperature and relative humidity measurement
- The controller reads data from all the sensors periodically at a frequency of 30 seconds and offloads it to ThingSpeak (a cloud based server employing MQTTS)

IoT Network Setup



- Location of the nodes and the traffic status of the roads (on an average day).
- These locations are used to collect the real-time traffic data as well as sensor data.
- These locations mainly contain major city roads and include a mixture of heavy and light traffic.
- The total distance covered is approximately 15 kms spanning an area of 6 km^2.

Dataset Collection 

- A dataset is collected using the PM monitoring node defined with the help of digital map service providers.

- A 5-dimensional feature vector has been accumulated for each data point in the dataset:
  - Traffic Mobility Rate (Using HERE Maps, Categorized into Free, Sluggish, Slow)
  - Normalized Difference Vegetation Index (NDVI) - Ranges from -1 (Water) to +1 (Vegetation)
  - Humidity - From PM Node
  - Temperature- From PM Node
  - Time of the day (categorized as morning, afternoon, and evening)

- After concatenating all the features accumulated from the samples in the dataset, a *m x 5* data matrix ***M*** is obtained, where *m* is the number of samples present in the dataset.

- A *m x 1* sized vector *y* containing the corresponding label for each sample is the respective AQI category computed using $PM_{2.5}$ and PM_{10} values.

Dataset Collection

AQI Categorization

- Each sampled data point of the dataset is associated with co-located respective node sensor values, i.e., temperature, relative humidity, $PM_{2.5}$, and PM_{10} measurement.

- The AQI level is computed using the $PM_{2.5}$, and PM_{10} values as per the Central Pollution Control Board, India[4], and categorized into five classes which are as follows:
  - Good (0 - 50)
  - Satisfactory (51 - 100)
  - Moderate (101 - 200)
  - Poor (201 - 300)
  - Severe (> 300)

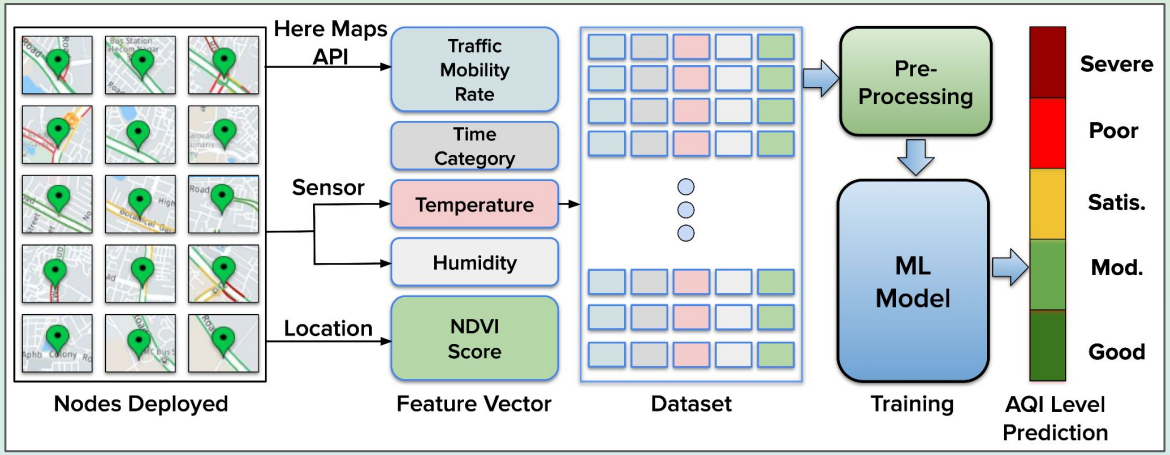
TMR	NDVI	Temperature	Humidity	Time of the Day
0.32	0.29	0.19	0.11	0.09

Importance of features w.r.t AQI

It can be observed that the feature **traffic mobility rate** and **NDVI score** plays an essential role with the support of temperature and rest other features.

4. National Air Quality Index, https://app.cpcbccr.com/AQI_India/.

Proposed Method



Algorithmic pipeline for the proposed methodology.

- Data Preprocessing
  - Standard Normalization
  - MinMax Scaler
- ML Model Training
  - Random Forest
  - Support Vector Machine
  - Multi Layer Perceptron
- ML Model Validation
  - Accuracy
  - Precision
  - Recall
  - F1-score
- Detection

Experimental Results

The ML models were trained and validated for the collected dataset on a total of 15 different nodes at different locations. As the environmental factor around them is diverse, two types of ML models were trained:

1. ML model on the overall dataset
2. Individual ML models for each node's dataset.

ML Model	Accuracy	Precision	Recall	F1-Score
RF	**82.60%**	**84.73%**	**82.63%**	**83.67%**
MLP	79.31%	77.98%	79.43%	78.70%
SVM	78.52%	77.13%	78.67%	77.89%

Performance of the various ML models on the overall dataset.

For the overall dataset, the RF model performed the best with **an accuracy of 82.6% and an F1-Score of 83.67%**

Experimental Results

# Node	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	79.69	78.16	79.93	79.03
2	79.56	78.11	81.03	79.54
3	78.21	79.55	78.26	78.90
4	78.51	79.36	79.61	79.48
5	79.86	78.75	78.42	78.58
6	82.78	84.70	81.72	83.18
7	78.95	79.90	79.3	79.59
8	80.15	82.42	81.98	82.20
9	79.56	79.96	79.90	79.92
10	78.98	82.49	80.29	81.37
11	81.67	84.60	81.26	82.89
12	82.94	84.31	80.82	82.52
13	79.88	84.78	79.30	81.94
14	81.05	83.44	80.37	81.87
15	82.63	79.96	78.65	79.29

Performance of the various ML models on the individual node dataset. (Best performing ML model result is shown)

For the individual node dataset, Node 6, 8, and 11 are near high vegetation areas. For these nodes, the AQI level for most of the data points fell in the first two categories, i.e., “Good” and “Satisfactory”. Hence, the model's task was easier for these nodes and performed better than the rest of the node's data.

On the other hand, for Node 1 and 3, the traffic mobility rate is high as they are placed at road junctions. For these nodes, the TMR varied mainly from “Sluggish” to “Slow”, resulting in Poor to Moderate AQI levels with some instances of Severe as well.

Conclusion 

- This paper introduced an IoT-based technique to predict the AQI from traffic and location data in real-time. Location-based features like traffic mobility rate, NDVI score, and sensor-based features like temperature and relative humidity were used to train the ML model.

- Additionally, a dataset having around 210,000 samples that contain traffic and weather information is collected.

- Experimental results show an **F1-Score of 83.67%** for the overall dataset, while experiments on node-specific datasets show the sensitiveness of the location.

- ML model performance on locations having high vegetation index performs better than others, specifically where the vegetation is low, and traffic is peak.

Thanks! 🙌