Fine-Grained Spatio-Temporal Particulate Matter Dataset From Delhi For ML based Modeling

Sachin K Chauhan, Sayan Ranu, Rijurekha Sen
Department of Computer Science
IIT Delhi
{csz188012, sayanranu, riju}@cse.iitd.ac.in

Zeel B Patel, Nipun Batra
Department of Computer Science
IIT Gandhinagar
{patel_zeel, nipun.batra}@iitgn.ac.in

Abstract

Air pollution poses serious health concerns in developing countries, such as India, necessitating large-scale measurement for correlation analysis, policy recommendations, and informed decision-making. However, fine-grained data collection is costly. Specifically, static sensors for pollution measurement cost several thousand dollars per unit, leading to inadequate deployment and coverage. To complement the existing sparse static sensor network, we propose a mobile sensor network utilizing lower-cost $PM_{2.5}$ sensors mounted on public buses in the Delhi-NCR region of India. Through this exercise, we introduce a novel dataset comprising $PM_{2.5}$ and PM_{10} measurements. This dataset is made publicly available at *https://www.cse.iitd.ac.in/pollutiondata*, serving as a valuable resource for machine learning (ML) researchers and environmentalists. We present two key contributions with the release of this dataset. Firstly, through in-depth statistical analysis, we demonstrate that the released dataset significantly differs from existing pollution datasets, highlighting its uniqueness and potential for new insights. Secondly, we conduct a benchmarking exercise (*https://github.com/sachin-iitd/DelhiPMDatasetBenchmark*), evaluating state-of-the-art methods for interpolation, feature imputation, and forecasting on this dataset, which is the largest publicly available PM dataset to date. The results of the benchmarking exercise underscore the substantial disparities in accuracy between the proposed dataset and other publicly available datasets. This finding highlights the complexity and richness of our dataset, emphasizing its value for advancing research in the field of air pollution.

1 Introduction

Air pollution has reached life-threatening levels in Delhi-National Capital Region (NCR), India [Tripathi *et al.*, 2019; Mannucci and Franchini, 2017], which is one of the most densely populated urban centers. The population of Delhi-NCR exceeds 46 million people [Nagar *et al.*, 2017] and it has been reported that 50% of all children staying in this region suffer from irreversible lung damage [Chatterji, 2021; ORF, 2021]. *Particulate Matter (PM)* is especially dangerous, since our breathing cannot filter out the ultra-fine particles. To mitigate the effects of air pollution, there is an urgent need to identify causes of pollution and strategies to curb its spread. It is suggested Sahu *et al.* [2020]; Sutaria [2022]

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

to use one sensor per km^2 for better pollution analysis. The *Central Pollution Control Board (CPCB)* and *Delhi Pollution Control Committee (DPCC)* have only 81 realtime air pollution measurement centers in Delhi-NCR Sutaria [2022] along with 65 manually monitored centers, which are thoroughly inadequate Guttikunda *et al.* [2023]; ET [2022] to cover the vast geography of $55,000$ km^2 NCRPB [2018].

In the literature, several models have been proposed for predicting pollution levels at same/future time points [Patel *et al.*, 2022; Gao and Li, 2021; Kurt *et al.*, 2008; Tsai *et al.*, 2018; Le *et al.*, 2020], and identifying factors affecting pollution [Apte *et al.*, 2011; Google, 2014; Messier *et al.*, 2018; Apte *et al.*, 2017; Alexeeff *et al.*, 2018]. There exists *interpolation models* [Qiao *et al.*, 2019; Rasmussen and Williams, 2005; Hamilton *et al.*, 2017; Patel *et al.*, 2022] to reliably predict pollution levels at unseen locations based on a sufficient number of pre-installed sensors. These models can improve with fine-grained pollution data. The interpolation and forecasting models are *supervised* in nature and hence can do better with more training data. Unfortunately, collecting pollution data using realtime centers is highly expensive as each instrument costs thousands of US Dollars.

In this work, we aim to mitigate the problem of lack of sufficient data in a cost-effective manner. We design a low-cost sensing mechanism (thoroughly compared in quality against high cost sensors) that allows us to collect PM data over a subset of the Delhi-NCR region at a fine spatio-temporal granularity. The key highlights and contributions of our work are:
1. Quality dataset: As it is not cost-effective to repeat even the low cost sensors per km^2, we establish a low-cost vehicle-mounted PM sensing network and release the largest $PM_{2.5}$ dataset from one of the most polluted regions in the world. This dataset is shown to be as good as the data collected from the few high-cost static-sensor deployed in the same region. As it is very challenging to collect such dataset in a developing country due to constraints in infrastructure and government permissions, we document our data collection experience briefly in the paper. (§ 3.2).
2. Unique dataset: This dataset complements the static sensor data available from the government deployed instruments in important ways. The static sensors are located at the top of high towers to get precise recordings of ambient pollution values, not affected by local sources. Our mobile sensors, on the other hand, are installed in the bus driver's cabin to measures the ground level pollution that daily commuters breathe in. We also perform a thorough comparison with PM datasets available from other parts of the world and establish that the released dataset is unique in terms of scale and statistical characteristics. Hence, it can be of immense value to environmental think tanks. (§ 3.3).
3. Utility for ML modeling: Through extensive benchmarking using state-of-the-art Machine Learning (ML) algorithms, we demonstrate the utility of this new dataset for modeling problems using ML, like spatio-temporal interpolation, missing data imputation and forecasting. The dataset is shown to be more challenging to model with ML algorithms, compared to previously available datasets, as Delhi has much higher variance in PM across space and time. This dataset, therefore opens opportunities for ML researchers for designing and benchmarking new ML algorithms, to reduce the interpolation, missing data imputation or forecasting errors. (§ 4).

2 Related Work

Spatio-temporal (ST) interpolation involves predicting air quality at unmonitored locations in the past and/or present time using training data observed from the sensors during the past and present time. Zheng *et al.* [2013] developed a co-training-based approach for ST interpolation using $PM_{2.5}$ values captured every hour from ground stations of 4 cities in China which are converted to AQI (Air Quality Index), along with meteorological and traffic data. Cheng *et al.* [2018] proposed an attention-based hybrid model involving LSTM and dense layers and Patel *et al.* [2022] proposed a domain-inspired non-stationary Gaussian process model for ST interpolation which can also be used for ST forecasting. The two used 36 monitoring stations in Beijing with the collection time interval of 1 hour (with the latter additionally using London data), alongside meteorological data.

Missing data imputation problem can be considered a variation of spatio-temporal interpolation where observations on the spatio-temporal cube are missing at random and we want to impute the missing data. Models that work for ST interpolation can mostly be adapted readily for this problem.

Spatio-temporal forecasting aims to predict air quality at a particular location in future using the past and current data available at all the installed sensors. Kurt *et al.* [2008] developed an online neural network based approach to predict air quality maximum 3 days ahead in time using 1 year PM_{10} data for 1 region in Turkey. Zheng *et al.* [2015] develop and deploy a machine learning based air quality forecasting system with the Chinese Ministry of Environmental Protection. Yi *et al.* [2018] develop a deep learning based approach to provide short-term, long-term air quality forecasts. The two used meteorological data along with pollution data generated every hour from 2,296 stations in 302 Chinese cities, and converted these concentrations into corresponding (individual) AQIs according to Chinese AQI standards. Air quality forecasting was posed as a challenge in KDD2018, where Luo *et al.* [2019] presented a winning solution based on a combination of classical machine learning and deep learning models using the provided data from stations in Beijing and London. Gao and Li [2021] propose a graph-based LSTM model for air quality forecasting and evaluate on Northwest China hourly data from 32 china stations.

All these prior arts utilize the static ground stations Air Quality data for the analysis, which enforces a restricted spatial coverage. They also use meteorological data from the respective regions. There also have been studies on low cost sensors available in market for developed (EU) regions Karagulian *et al.* [2019] only. Also, a project about installing low cost sensors at different roadside locations Schneider *et al.* [2023] to complement the existing expensive static sensor network is done recently, but they kept the sensors at fixed locations. We are working on the PM data collected with mobile sensors, which is fine-grained and provides better spatio-temporal coverage, and our benchmarked models do not rely on other meteorological factors.

3 Dataset Description

3.1 Dataset Collection Challenges

Creating the mobile PM dataset (as a replacement for low cost static PM dataset and high-cost ground station PM dataset) required us to design and implement our own embedded platform, choosing and calibrating appropriate sensors for maximum accuracy at low cost. We opted to install our device in public buses, to utilize their pre-defined/fixed and frequent routes of travel. Packaging was challenging to securely mount the instruments in the public buses, avoiding theft and ensuring enough ambient air to measure PM. Cellular connectivity was intermittent as the buses traversed the city, requiring us to augment real time data transfer when signal was present, with local storage to save data when signal strength dropped. Finally, getting permissions from different government entities to instrument the public bus fleet needed strict safety certifications that our devices do not interfere with the electrical and mechanical functioning of the bus.

We mounted pollution tracking sensors on the permissible 13 public buses in Delhi for 3 months (Nov 1^{st}, 2020 to Jan 31^{st}, 2021), in collaboration with Delhi Integrated Multimodal Transport System, after rigorous tests for automotive safety certification and appropriate permissions and letters of support from the Delhi Ministry of Transport and Delhi Pollution Control Committee. The inside of our custom-made instrument comprising *(a)* PM sensor measuring $\text{PM}_{2.5}$, PM_{10} and PM_1, *(b)* GPS sensor to locate the bus, *(c)* 4G radio to communicate data from bus to server, *(d)* SD card for locally storing data when 4G signal is unavailable, *(e)* BME sensor BME [2023], a sensor especially developed for mobile applications and wearables, to record temperature and relative humidity and *(f)*

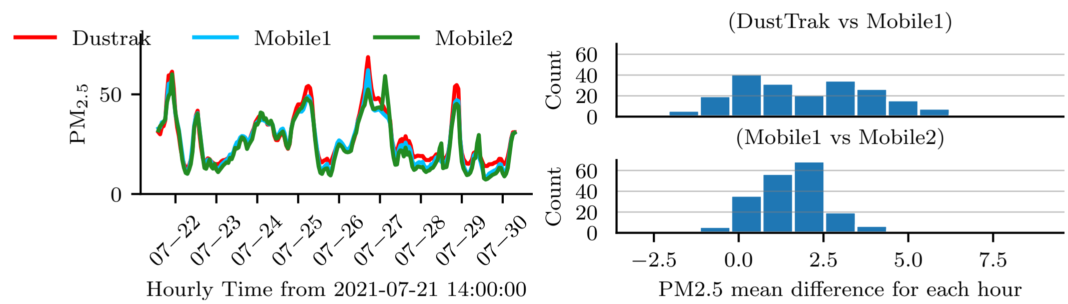
(a) Measuring device (b) Mounting location (c) Mounted device (d) Bus trajectories

Figure 1: (a) Inside of our PM measuring IoT unit. (b) Mounting location in bus driver's cabin in non air-conditioned public bus (below the existing white box). (c) Mounted IoT unit in the bus (below the existing white box). (d) Government deployed static sensors installed in and around our bus trajectories, as location icons.

micro-controller to orchestrate the sense-store-communicate software (See Fig. 1a). The mounting location in the bus driver's cabin, next to two open windows to allow enough air-flow (Fig. 1b-1c). Each bus commutes for 16-20 hours per day, and our instruments collect data at a fine granularity of 20 samples per minute. Overall, the bus trajectories cover 559 km^2, along the main arterial roads in North-West, North, North-East and South-East Delhi (Fig. 1d). The dataset has been made available at *https://www.cse.iitd.ac.in/pollutiondata/* with proper documentation, under a Creative Commons Attribution 4.0 International License CC-by4 [2013].

3.2 Data Quality Analysis

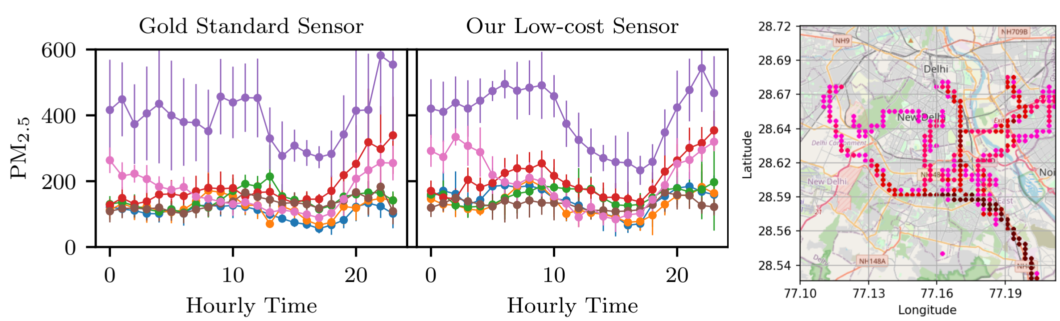
Fig. 2a plots $PM_{2.5}$ values measured by two low cost PM sensors built by us (cost USD 30), and the same measured by an industry grade reference instrument TSI DustTrak (cost USD 9500), while all three instruments are placed close to each other. The plot shows hours of the day along x-axis and sensed $PM_{2.5}$ values along y-axis, for 10 sample days Jul 21-31, 2021. This is after the deployment of the low cost sensors in the buses is over, and the sensors have been brought back to the lab. Fig. 2b shows the histogram of difference of hourly mean PM between DustTrak and one mobile sensor, and two low cost mobile sensors, for the same 10 days. While the cost gap between the instruments is huge, the gap between their sensed $PM_{2.5}$ values, as seen in this graph, is negligible. This pattern has been observed consistently by us and other researchers [Zheng *et al.*, 2018; Cheng *et al.*, 2014; Gao *et al.*, 2015; Rai *et al.*, 2017; Jiao *et al.*, 2016; Zheng *et al.*, 2019].



(a) Time-series comparison (b) Histogram of differences

Figure 2: (a) $PM_{2.5}$ values measured by our low-cost mobile PM sensors (USD 30) vs. TSI DustTrak (USD 9500) between Jul 21-31, 2021. (b) Histogram of pointwise differences of $PM_{2.5}$ values measured by DustTrak and low cost mobile PM sensors. The values are almost identical.

We additionally compare the distribution of PM values recorded by our mobile sensors vs. those by the high-cost static sensors, deployed at sparse locations by CPCB and DPCC in Delhi-NCR. Fig. 3a(Left) shows hours of day along x-axis and average $PM_{2.5}$ for that hour, as measured by reference grade static monitors, with standard-deviation bars along y-axis. Fig. 3a(Right) shows the same averaged over all bus mounted sensors. We select the static sensors that are within 1km of mobile sensor trajectory for each hour, and plot for 7 sample days. Fig. 3a reveal that both static and bus mounted sensors show similar PM distributions for each day, in spite of the difference in heights



(a) Static v/s Mobile Sensors Data (b) Heat Map

Figure 3: (a) Distribution of $PM_{2.5}$ collected by our low-cost sensor and gold standard sensor over 7 random days. The distributions are similar across the two sets of instruments. (b) Heat Maps (darker locations contain more samples).

they have been installed at, and the difference in PM measurement technique. We see this agreement for the entire 3 months deployment period. The agreement between low cost mobile sensors, and a co-located high cost TSI Dusttrak, as well as reference grade static monitors, give us confidence to release the dataset to the research community.

Heat Map: During our analysis, we discovered variations in data availability across different timestamps and spatial locations. It was evident that certain timestamps were not available at all spatial locations. Furthermore, some spatial locations, which were situated along routes with fewer bus visits throughout the day, exhibited limited temporal samples. As illustrated in Fig. 3b, a typical day (Dec 15, 2020) demonstrated this pattern, where the outermost locations (depicted in light/pink color) contained samples from 4 hours duration within the 16.5-hour effective temporal window. Conversely, the darker/brown locations near the bottom right of the figure displayed a higher number of samples, ranging from 14 to 16.5 hours. These locations are associated with common bus routes that connect with the depot.

3.3 Dataset Novelty

Tables 1 and 2 summarize the statistics of the dataset. While vehicle mounted air pollution sensing has been conducted [Apte *et al.*, 2011; Google, 2014; Apte *et al.*, 2017; Alexeeff *et al.*, 2018; Guo *et al.*, 2016; Adams and Corr, 2019; Li *et al.*, 2012], our dataset is unique in characteristics and scale. Specifically, only two studies from Ontario, Canada [Adams and Corr, 2019] and Zurich, Switzerland [Li *et al.*, 2012] have made their datasets publicly available. The Zurich dataset does not include PM values. Compared to the Canada dataset, our dataset is 1000 times larger and has a significantly different distribution of PM values (See Tables 1 and 2). This is understandable as Delhi-NCR is an air pollution hotspot, whereas Zurich and Ontario have negligible PM levels. We also compare our dataset with a recent USA AQI dataset Bhattacharyya *et al.* [2022] collected from Air Quality Open Data Platform.

Table 1: Details of Delhi, India and Hamilton, Ontario, Canada and USA datasets.

Metric	Delhi-NCR	Canada	USA
Total area	559 km^2	1138 km^2	54 cities
Total samples	12,542,183	46,080	35,596
Samples with PM2.5	12,542,183	12,154	35,134
Pollutants covered	PM_1, $PM_{2.5}$ and PM_{10}	CO, NO, NO_2, SO_2, O_3, PM_1, $PM_{2.5}$ and PM_{10}	CO, NO_2, SO_2, O_3, $PM_{2.5}$ and PM_{10}
Sensor source	Public bus	Commercial van	OpenDataPlatform
Monitoring days	91	114	668

Table 2: Statistical comparison of PM values in Delhi, Canada and USA datasets.

Metric	Delhi-NCR			Canada			USA	
	PM_1	$PM_{2.5}$	PM_{10}	PM_1	$PM_{2.5}$	PM_{10}	$PM_{2.5}$ AQI	PM_{10} AQI
Mean	120.35	207.92	226.11	12.15	15.08	46.45	31.15	17.67
Std-dev	57.27	114.36	123.86	9.02	12.87	97.36	17.11	11.00
Missing %	0	0	0	71.71	73.62	72.24	1.30	52.34

4 ML Modeling Benchmarks

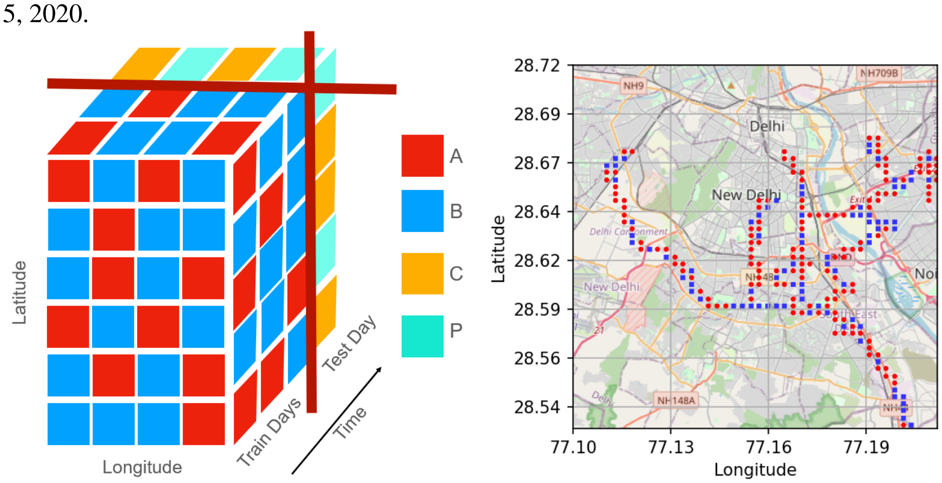
In this section, we benchmark the machine learning problems of **(1)** spatio-temporal interpolation, **(2)** spatio-temporal data imputation and **(3)** spatio-temporal forecasting on the proposed and the Canada datasets. This benchmarking study serves two roles. First, it allows us to compare the complexities of the two datasets beyond just statistical characterization. Secondly, spatio-temporal interpolations, data imputations, and forecasting methods are crucial for environmental research, policy-making, and individual decision-making. They empower various stakeholders to gain a comprehensive understanding of air pollution, proactively address potential increases in pollution levels, and make informed choices to reduce personal exposure. In order to harness the full potential of spatio-temporal forecasting, interpolations, and data imputations, it is crucial to benchmark and evaluate the performance of algorithms designed to tackle these problems.

4.1 Dataset Pre-processing and Evaluation Metrics for the Analysis

To benchmark ML modeling algorithms, we process and split the data into two parts for *visible* and *held-out/hidden*. For the Delhi dataset, we focus on the data collected from Nov 12, 2020, to Jan 30, 2021, excluding the initial days when there were fewer instruments on the buses and limited sample data. Additionally, we exclude the nightly data between 10 PM IST and 5:30 AM IST when buses remain stationary at a confined bus-depot. To facilitate analysis, we divide the geographical area into square spatial grids with a side length of 1 km. These grids are further converted into spatio-temporal cells with a time interval of 30 minutes. To obtain representative PM values, we compute the average of all samples within each spatio-temporal cell. Subsequently, we employ K-fold cross-validation to partition the data into K PM visible / held-out sets for each day. The results obtained from the Delhi dataset are denoted as *Delhi (Day)* in the generated plots.

Additionally, we utilize two open-sources PM datasets, from Hamilton in Ontario, Canada Adams and Corr [2019] and from USA Bhattacharyya *et al.* [2022]. For the Canada dataset, we process the data from 18 distinct days in the year 2015 using the same methodology. These results are presented as *Canada (Day)* in the respective experiments. As the data for Canada exhibits temporal sparsity, we project the data for each year onto a single day and treat it as equivalent to 11 days (from 2006 to 2016). The outcomes of this processing approach are depicted as *Canada (Year)* in the experiments. For the USA data, we use the available PM data across 54 cities from Jan 1, 2019 to Dec 11, 2020, and the results are presented as *USA (Day)*. We benchmark the datasets on Nvidia DGX Workstation (with 4X Tesla V100 GPUs) and the benchmarking code is available at *https://github.com/sachin-iitd/DelhiPMDatasetBenchmark*.

Notation: We use T (consecutive) days data for the training and take the next day for test/evaluation. Fig. 4a denotes the various subsets of this T+1 days data as A, B, C and P. For a given fold, A is the visible set with 80% of all T train days data, B is the held-out set with the remaining 20% of the T train days data. $\mathrm{A} \cup \mathrm{B}$ forms the whole dataset for the T train days. C is the visible set with 80% of the test day data, P is the held-out set with the remaining 20% of the test day and $\mathrm{C} \cup \mathrm{P}$ forms the whole dataset for the test day. The exact number of locations in A, B, C and P change across the K folds. In Fig. 4b, we show set of A and B spatial locations in Delhi dataset for 3 PM to 4:30 PM on Dec 15, 2020.



(a) A, B, C and P sets. (b) Visible (A) and Held-out (B) sets over map.

Figure 4: PM Data Splits.

4.2 Formulation of different ML Prediction Problems

(a) Spatio-temporal Interpolation: Given set of visible locations A and C where we have input features (latitude, longitude and time) and $\mathrm{PM}_{2.5}$ available for T+1 days, we wish to estimate $\mathrm{PM}_{2.5}$ for a set of held-out locations P for the T+1th day using the input features (latitude, longitude and time). This approach is compatible to the scenario where we have data for some locations and we use interpolation algorithms to know the PM values at new locations.

The Loss is computed as follows:

$$RMSE(L'_p, L_p) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y'_i - y_i)^2} \tag{1}$$

where y'_i is the predicted and y_i is the true $PM_{2.5}$ value, and N is the total number of samples.

For each of the K folds, we separately compute RMSE of prediction over P for that fold, and then plot average with standard deviation bars over the K folds. The lower RMSE being the better.

(b) Spatio-temporal Missing Data Imputation: Given set of visible locations A and C where we have input features (latitude, longitude and time) and $PM_{2.5}$ available for T+1 days and a set of held-out locations B where we have input features (latitude, longitude and time) and $PM_{2.5}$ available for T days, we wish to estimate $PM_{2.5}$ for a set of held-out locations P for the T+1th day using the input features (latitude, longitude and time). This setting is compatible to the scenario where we have intermittent data missing throughout the day and we use interpolation algorithms to predict the missing points taking past and present data as input.

(c) Spatio-temporal Forecasting: Given a set of locations A and B where we have input features (latitude, longitude and time) and $PM_{2.5}$ available for T days, we wish to estimate $PM_{2.5}$ for a set of locations C and P for the T+1th day using the input features (latitude, longitude and time). As all the data is involved in training and evaluation, different splits from the K-fold are not required.

4.3 ML Algorithms Benchmarked in this Paper

(a) Mean Predictor is the simple mean value of all visible samples is used as the value of the held-out locations. The mean value of all visible $PM_{2.5}$ locations C is used as the value of the held-out $PM_{2.5}$ locations P.

$mean \leftarrow \frac{1}{|C|}\sum PM_{2.5}^{c} \;\; \forall c \in C$

$PM_{2.5}^{p} \leftarrow mean \;\; \forall p \in P$

(b) Inverse Distance Weighting (IDW) is the weighted average value of all visible C samples in terms of distance, is used as the value of the held-out P locations.

$PM_{2.5}^{p} \leftarrow \sum \frac{PM_{2.5}^{c}}{F(d_{cp})} \;\; \forall c \in C \;\; \forall p \in P$

where F is a linear function on distance d.

(c) Random Forest (RF) is a non-linear model capable of modeling complex spaces. It is known to perform efficiently on non-linear regression tasks, using an ensemble of multiple decision trees, taking the final output as the mean of the output from all trees.

(d) XGBoost (XGB) iteratively combines the results from weak estimators. It uses gradient descent while adding new trees during training.

(e) ARIMA or Auto-Regressive Integrated Moving Average is a statistical time-series forecasting model that uses linear regression. It is configured using parameters (p, d, q) as: p is the number of lag observations included in the model, d is the number of times raw observations are differenced, and q is the size of the moving average window. We use ARIMA with parameters (3, 1, 1).

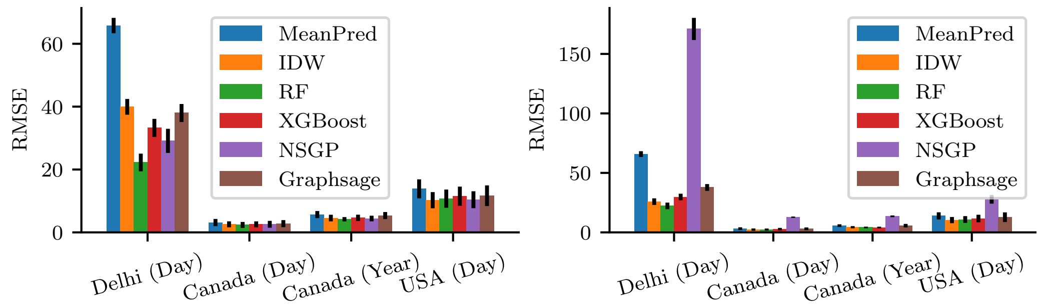
(f) N-BEATS is Neural Basis Expansion Analysis for Time Series, a deep learning model for zero-shot time-series forecasting Oreshkin *et al.* [2020]. We use the code from Python library "Darts".

(g) Non-Stationary Gaussian Process (NSGP) is a gaussian processes based baseline taken from AAAI 2022 Patel *et al.* [2022]. It learns a non-stationary covariance Plagemann *et al.* [2008] for latitude and longitude and locally periodic covariance for time. In general, Gaussian process a.k.a. Kriging is a Bayesian non-parametric model known as the best unbiased predictor in spatial interpolation domain Rasmussen and Williams [2005]. It conditions on the training data and provides a posterior predictive distribution at the new locations with closed form equations. With only three tunable parameters, it is considered a strong baseline in spatial interpolation domain.

(h) Graphsage is a graph neural network model to learn and predict values at unknown spatio-temporal locations Hamilton *et al.* [2017]. We transform the PM data to a graph, and use Graphsage for interpolation and missing data imputation. Our graph formulation is available in Appendix A.

4.4 Observations and Inferences

Fig. 5, shows the RMSE for interpolation, using 5-fold cross validation for the two training configurations ACT in Fig. 5a and C in Fig. 5b, for 3 training days. ACT uses the visible set from both training and test days, while C uses only the test day's PM visible set. The missing data imputation plots are almost identical to the interpolation plots, so we omit these for space constraints.



(a) Train:ACT, Input:ACT, Predict:P

(b) Train:C, Input:C, Predict:P

Figure 5: Interpolation RMSE. Training days' data is used by ML model in (a) and not used in (b).

Observation 1: Delhi dataset is harder to model. All experiments over Delhi data show higher RMSE and all experiments over Canada and USA data show low RMSE, for both interpolation and forecasting, in Figures 5, and 6. This shows that Delhi data is more challenging for ML modeling, than the currently available PM datasets.

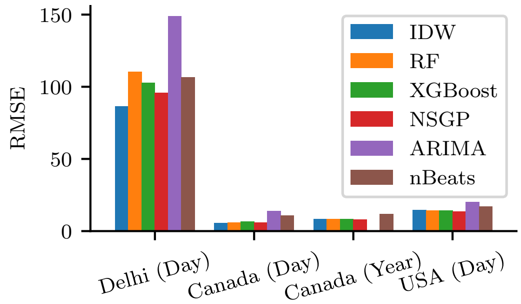
Observation 2: Learning from data helps in modeling the Delhi dataset. All ML based algorithms show significant improvement over Mean Predictor for Delhi data in Figures 5a, whereas improvement for Canada and USA data over Mean Predictor is not significant. In Fig. 5a, all ML algorithms exhibit less than 40 RMSE while Mean Predictor RMSE is 65.80 for Delhi data (best case improvement is 66.2% for RF and worst case 39.3% for IDW). For Canada data, best case improvement is $\sim$ 27% and worst case sees no improvement, whereas for USA AQI data, improvement is within 16% - 26%.

Observation 3: Traditional ML algorithms do as well as the recent models for the Delhi dataset. Learning from data matters, as the ML based models do better than the mean predictor. But the recent complex Bayesian models like NSGP, and the neural network based models like Graphsage (for interpolation) and N-BEATS (for forecasting), do not outperform powerful traditional ML models like Random Forest. For instance, RF performs best for interpolation (RMSE 22.24 in Fig. 5), and XGBoost performs best for forecasting (RMSE 84.15 in Fig. 6).

Observation 4: Historical training data adds no value for interpolation. For the spatio-temporal interpolation problem, just using data from the visible set C from test day is enough to predict the held-out P data with low RMSE. For example, the RMSE for RF is similar (22.24) for test day only data C in Fig. 5b and with including train day data ACT in Fig. 5a. And XGBoost is better for C with RMSE 29.73 than for ACT with RMSE 33.24. NSGP is the only algorithm, which sees a huge jump in RMSE when not using training data from past days. Thus PM for a given day is mostly unrelated to PM on past days, and using historical training data has no significant impact on interpolation RMSE.

Fig. 6 shows RMSE of forecasting. Graphsage does not work in this setting as it requires a subset of test day's data for edge formation to the data being predicted. So we drop Graphsage, and add two forecasting specific baselines: ARIMA and N-BEATS, that are not suitable for interpolation.

Observation 5: Forecasting is a harder problem than interpolation. Forecasting RMSEs are significantly higher than interpolation RMSEs. The best model in forecasting is XGBoost in Fig. 6 with RMSE 84.15, whereas the best model for interpolation in Fig. 5 is RF with RMSE 22.24. Higher forecasting RMSE compared to interpolation also supports that previous day's data has less impact on test day's PM data. Hence forecasting using only past days' data for an unseen future test day is hard.



(a)

	Config		Delhi (Day)	
Algo	Train	Input	Train	Test
IDW	ABT	ABT	-	86.52
	AB	AB	-	270.73
RF	ABT	ABT	8.09	110.49
	AB	AB	45.85	89.54
XGBoost	ABT	ABT	33.19	102.68
	AB	AB	57.12	84.15
NSGP	ABT	ABT	24.34	95.83
	AB	AB	51.90	86.34
ARIMA	ABT	ABT	204.87	148.86
N-BEATS	ABT	ABT	-	106.41

(b)

Figure 6: Forecasting RMSE. (a) Normalization is done across all days. (b) Comparison of normalization across all days (T) vs normalization over each day.

Observation 6: How time is normalized affects forecasting accuracy. In Fig. 6a, time normalization is done across days, i.e. time starts at 0 on first train day and increases to 1 till last train day. ARIMA / N-BEATS don't normalize the time directly, they take all PM values in a sequence corresponding to time from start to end. RF/XGBoost takes input in random sequence and hence takes the time as a state parameter, which can be normalized from start to end, or for each day. Table 6b compares this time normalization across days (T), to normalizing separately for each day. RF, XGBoost and NSGP show lower RMSE for separate normalization for each day, while IDW does better with normalization across days. This pre-processing step of time normalization therefore should be carefully decided based on the ML algorithm.

5 Conclusion and Future Work

Delhi-NCR, with its notorious air pollution problem, poses a significant health risk to its population of approximately 46 million individuals. In this paper, we present a novel PM dataset collected from this region using low-cost IoT devices deployed on public buses. This dataset serves as a valuable resource for environmental researchers and medical practitioners, offering insights into ground-level PM exposure for daily commuters and temporal variations in PM levels over days and weeks. Moreover, it provides a comprehensive view of spatial variations across different locations within the region.

Through thorough statistical analysis and benchmarking studies, we have established that the released dataset is distinct from any other existing pollution dataset. By comparing the performance of machine learning algorithms on the released dataset against the Canada dataset, we have demonstrated the significant differences in characteristics and challenges associated with the Delhi-NCR dataset. This highlights the need for specialized approaches and tailored solutions to address the unique complexities of air pollution in this region.

The availability of this low-cost mobile monitoring system has the potential to complement the expensive static sensor network in the city, empowering citizens to make informed decisions regarding local PM levels. This includes determining the safety of engaging in outdoor activities, choosing appropriate protective measures such as face-masks or air purifiers, and selecting optimal commuting routes and transportation modes to minimize PM exposure. Such considerations are vital for safeguarding public health and promoting environmental sustainability.

In our future work, we aim to address the problem of recommending suitable locations for installing new expensive sensors effectively within budget constraints, a challenging task in a developing country like India. By leveraging the insights gained from this research, we strive to optimize the allocation of resources and enhance the efficiency of the monitoring network, further strengthening pollution mitigation efforts. To foster further advancements in the field of environmental sustainability, we release both the code and data associated with this study. This allows researchers to build upon our work, explore new avenues of inquiry, and contribute to the collective understanding and management of air pollution-related challenges.

References

Matthew D. Adams and Denis Corr. A mobile air pollution monitoring data set. *Data*, 4(1), 2019.

Stacey E Alexeeff, Ananya Roy, Jun Shan, Xi Liu, Kyle Messier, Joshua S Apte, Christopher Portier, Stephen Sidney, and Stephen K Van Den Eeden. High-resolution mapping of traffic related air pollution with google street view cars and incidence of cardiovascular events within neighborhoods in oakland, ca. *Environmental Health*, 17:1–13, 2018.

Joshua S Apte, Thomas W Kirchstetter, Alexander H Reich, Shyam J Deshpande, Geetanjali Kaushik, Arvind Chel, Julian D Marshall, and William W Nazaroff. Concentrations of fine, ultrafine, and black carbon particles in auto-rickshaws in new delhi, india. *Atmospheric Environment*, 45(26):4470–4480, 2011.

Joshua S Apte, Kyle P Messier, Shahzad Gani, Michael Brauer, Thomas W Kirchstetter, Melissa M Lunden, Julian D Marshall, Christopher J Portier, Roel CH Vermeulen, and Steven P Hamburg. High-resolution air pollution mapping with google street view cars: exploiting big data. *Environmental science & technology*, 51(12):6999–7008, 2017.

Mayukh Bhattacharyya, Sayan Nag, and Udita Ghosh. Deciphering environmental air pollution with large scale city data. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5031–5037. International Joint Conferences on Artificial Intelligence Organization, 7 2022. AI for Good.

BME. Humidity sensor bme280, 2023.

CC-by4. Attribution 4.0 international (cc by 4.0), 2013.

Arpan Chatterji. Air pollution in delhi: filling the policy gaps. *Massach Undergr J Econ*, 17(1), 2021.

Yun Cheng, Xiucheng Li, Zhijun Li, Shouxu Jiang, Yilong Li, Ji Jia, and Xiaofan Jiang. Aircloud: A cloud-based air-quality monitoring system for everyone. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*, SenSys '14, 2014.

Weiyu Cheng, Yanyan Shen, Yanmin Zhu, and Linpeng Huang. A neural attention model for urban air quality inference: Learning the weights of monitoring stations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

ET. Caqm asks delhi ncr states to install sensors to check pollution at construction sites and hotspots, 2022.

Xi Gao and Weide Li. A graph-based lstm model for pm2. 5 forecasting. *Atmospheric Pollution Research*, 2021.

Meiling Gao, Junji Cao, and Edmund Seto. A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of pm2. 5 in xi'an, china. *Environmental pollution*, 199:56–65, 2015.

Google. Mapping the invisible: Street view cars add air pollution sensors, 2014.

Hongjie Guo, Guojun Dai, Jin Fan, Yifan Wu, Fangyao Shen, and Yidan Hu. A mobile sensing system for urban monitoring with adaptive resolution. *Journal of Sensors*, 2016, 2016.

Sarath K. Guttikunda, Sai Krishna Dammalapati, Gautam Pradhan, Bhargav Krishna, Hiren T. Jethva, and Puja Jawahar. What is polluting delhis air? a review from 1990 to 2022. *Sustainability*, 15(5), 2023.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *31st NeurIPS Conference*, 2017.

Wan Jiao, Gayle Hagler, Ronald Williams, Robert Sharpe, Ryan Brown, Daniel Garver, Robert Judge, Motria Caudill, Joshua Rickard, Michael Davis, et al. Community air sensor network (cairsense) project: evaluation of low-cost sensor performance in a suburban environment in the southeastern united states. *Atmospheric Measurement Techniques*, 9(11), 2016.

Federico Karagulian, Maurizio Barbiere, Alexander Kotsev, Laurent Spinelle, Michel Gerboles, Friedrich Lagler, Nathalie Redon, Sabine Crunaire, and Annette Borowiak. Review of the performance of low-cost sensors for air quality monitoring. *Atmosphere*, 10(9), 2019.

Atakan Kurt, Betul Gulbagci, Ferhat Karaca, and Omar Alagha. An online air pollution forecasting system using neural networks. *Environment international*, 2008.

Van-Duc Le, Tien-Cuong Bui, and Sang-Kyun Cha. Spatiotemporal deep learning model for citywide air pollution interpolation and prediction. In *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 55–62. IEEE, 2020.

Jason Jingshi Li, Boi Faltings, Olga Saukh, David Hasenfratz, and Jan Beutel. Sensing the air we breathe: The opensense zurich dataset. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, page 323–325. AAAI Press, 2012.

Zhipeng Luo, Jianqiang Huang, Ke Hu, Xue Li, and Peng Zhang. Accuair: Winning solution to air quality prediction for kdd cup 2018. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1842–1850, 2019.

Pier Mannuccio Mannucci and Massimo Franchini. Health effects of ambient air pollution in developing countries. *International journal of environmental research and public health*, 14(9):1048, 2017.

Kyle P Messier, Sarah E Chambliss, Shahzad Gani, Ramon Alvarez, Michael Brauer, Jonathan J Choi, Steven P Hamburg, Jules Kerckhoffs, Brian LaFranchi, Melissa M Lunden, et al. Mapping air pollution with google street view cars: Efficient approaches with mobile monitoring and land use regression. *Environmental science & technology*, 52(21):12563–12572, 2018.

Pavan K Nagar, Dhirendra Singh, Mukesh Sharma, Anil Kumar, Viney P Aneja, Mohan P George, Nigam Agarwal, and Sheo P Shukla. Characterization of pm 2.5 in delhi: role and impact of secondary aerosol, burning of biomass, and municipal solid waste and crustal matter. *Environmental Science and Pollution Research*, 24:25179–25189, 2017.

William Navidi. *Statistics for Engineers and Scientists*. McGraw-Hill, 2009.

NCRPB. Ncr constituent areas, 2018.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020.

ORF. Delhi is failing its children, air pollution is choking their future, 2021.

Zeel B Patel, Palak Purohit, Harsh M Patel, Shivam Sahni, and Nipun Batra. Accurate and scalable gaussian processes for fine-grained air quality inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12080–12088, Jun. 2022.

Christian Plagemann, Kristian Kersting, and Wolfram Burgard. Nonstationary gaussian process regression using point estimates of local smoothness. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II 19*, pages 204–219. Springer, 2008.

Pengwei Qiao, Peizhong Li, Yanjun Cheng, Wenxia Wei, Sucai Yang, Mei Lei, and Tongbin Chen. Comparison of common spatial interpolation methods for analyzing pollutant spatial distributions at contaminated sites. *Environmental geochemistry and health*, 41(6):2709–2730, 2019.

Aakash C Rai, Prashant Kumar, Francesco Pilla, Andreas N Skouloudis, Silvana Di Sabatino, Carlo Ratti, Ansar Yasar, and David Rickerby. End-user perspective of low-cost sensors for outdoor air pollution monitoring. *Science of The Total Environment*, 607:691–705, 2017.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

Ravi Sahu, Kuldeep Kumar Dixit, Suneeti Mishra, Purushottam Kumar, Ashutosh Kumar Shukla, Ronak Sutaria, Shashi Tiwari, and Sachchida Nand Tripathi. Validation of low-cost sensors in measuring real-time pm10 concentrations at two sites in delhi national capital region. *Sensors*, 20(5), 2020.

Philipp Schneider, Matthias Vogt, Rolf Haugen, Amirhossein Hassani, Nuria Castell, Franck R. Dauge, and Alena Bartonova. Deployment and evaluation of a network of open low-cost air quality sensor systems. *Atmosphere*, 14(3), 2023.

Howard Seltman. *Experimental Design and Analysis*. Carnegie Mellon University, 2018.

Ronak Sutaria. Delhi plans mesh of sensors to monitor pollution air hot spots, 2022.

CB Tripathi, Prashant Baredar, and Lata Tripathi. Air pollution in delhi. *Current Science*, 117(7):1153–1160, 2019.

Yi-Ting Tsai, Yu-Ren Zeng, and Yue-Shan Chang. Air pollution forecasting using rnn with lstm. In *2018 IEEE 16th Intl DASC/PiCom/DataCom/CyberSciTech Conf*. IEEE, 2018.

Xiuwen Yi, Junbo Zhang, Zhaoyuan Wang, Tianrui Li, and Yu Zheng. Deep distributed fusion network for air quality prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 965–973, 2018.

Yu Zheng, Furui Liu, and 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*, pages 1436–1444, 2013.

Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. Forecasting fine-grained air quality based on big data. In *Proceedings of the 21th SIGKDD conference on Knowledge Discovery and Data Mining*. KDD 2015, August 2015.

Tongshu Zheng, Michael H. Bergin, Karoline K. Johnson, Sachchida N. Tripathi, Shilpa Shirodkar, Matthew S. Landis, Ronak Sutaria, and David E. Carlson. Field evaluation of low-cost particulate matter sensors in high and low concentration environments. *Atmospheric Measurement Techniques*, 2018.

T. Zheng, M. H. Bergin, R. Sutaria, S. N. Tripathi, R. Caldow, and D. E. Carlson. Gaussian process regression model for dynamically calibrating and surveilling a wireless low-cost particulate matter sensor network in delhi. *Atmospheric Measurement Techniques*, 12(9):5161–5181, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [N/A] We use cost effective approaches with possible limitation in accurate sensing compared to the standard expensive instruments.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Refer § 3.1 and § 4.1.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Refer § 4.1.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] Using data open-sourced in previous research works giving appropriate reference.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] Used open-source or self-curated datasets.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Pollution Data does not contain such content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Graphsage (with Graph formulation)

We aim at learning universal weights, similar to GraphSAGE Hamilton *et al.* [2017], which will signify the importance of a neighbour based on some known node values and edge weights. Here we define node values as the value of the pollutant PM2.5 while the edges are created using latitude, longitude and datetime features. Firstly, a graph is created from the train dataset, aggregating all inputs within 500m and 30 minutes of each other into a single node. An edge is created between two nodes if they lie within 2 hours of each other. The graph then goes through two graph-based layers to learn the required weights where embeddings are learnt using the max and mean aggregation layers, followed by 3 fully connected neural network layers to predict the final pollutant value.

Let $G = (V, E, \sigma, \mathcal{A})$ be a Directed Graph with V vertices/nodes, E edges, $\mathcal{A}$ attributes and σ as the label mapping, where

$\sigma : V \rightarrow \mathcal{L}$

$\mathcal{L}$ being the set of $PM_{2.5}$ values.

V corresponds to the spatiotemporal locations where $PM_{2.5}$ values are known (S: Red) or desired (U: Blue), i.e. V=S+U. E ($e \in E$) connects the V ($v \in V$) such that

e_{ij} = (v_i, v_j) | $v_i \in S \wedge v_j \in (S \vee U)$ and $t_{ij} \leq TimeLimit$, where t_{ij}=abs(v_i^t - v_j^t)

The Graph G comprises of separate connected components for different days.

e_{ij} = (v_i, v_j) | $v_i \in Day_p$ and $v_j \in Day_q \Rightarrow p = q$

Weight of each edge is inversely proportional to the spatial distance between the two nodes across the edge.

$w_{ij} = \frac{1}{1+d_{ij}}$, if e_{ij} exists, where d_{ij}=haversine(v_i, v_j)

Edges exist from all S nodes to each U node. No S to S edges exist.

$e_{ij} = (v_i, v_j) \mid v_i \in S$ and $v_j \in U \Rightarrow |e_{ij}\ \forall i| = |S|\ \forall j$

The graph G is of two types:

Train Graph G_{Train}**:** It is used for training Graphsage Neural Network.

$v \in Day_{Train} \Rightarrow v \in S \vee U \Rightarrow |v \in S| > 0$ and $|v \in U| > 0$

The RMSE loss on the nodes $v \in U$ is used for model training.

Test Graph G_{Test}**:** It is used for evaluating the trained Graphsage model on unseen test day data (Day_{Test}) along with full data from known days.

$v \in Day_{Test} \Rightarrow v \in S \vee U \Rightarrow |v \in S| > 0$ and $|v \in U| > 0$

The v is formed by taking the corresponding $PM_{2.5}$ label L and an indicator variable I.

$v_i = L_i | I_i$

$L_i \leftarrow PM_{2.5}, I_i \leftarrow= 1\ \forall\ v \in S$

$L_i \leftarrow 0, I_i \leftarrow= 0\ \forall\ v \in U$

The 2 layer mean-pool and max-pool model graphsage architecture is shown in Fig. 7.

The RMSE loss of the nodes $v \in U$ (or $v \in P$ in particular) is used as the reporting metric.

For Graphsage based evaluation, out the 80% training data in 5-fold cross validation, we use 40% as *visible* set, 40% as *held-out* set, to manage edges between these two sets.

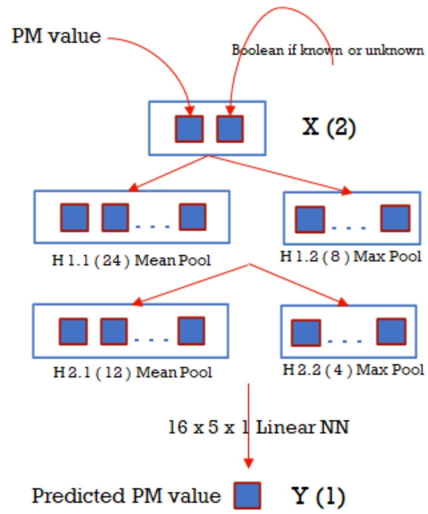


Figure 7: Graphsage model architecture.

B Complete ML Benchmarks

Table 3 shows the complete benchmark for Spatio-temporal Interpolation for different train and input configurations. An important subset of these benchmarks is presented in Fig. 5 and discussed in § 4.4 in the main paper. The benchmarks for NSGP algorithm for some configurations for USA dataset (marked by * in Table 3) is in progress and cannot be completed yet due to resource constraints, for which we present the partial results and mark accordingly.

Table 3: Spatiotemporal Interpolation RMSE for different configurations (* denotes partial experiments).

Algo	Config		Delhi (Day)		Canada (Day)		Canada (Year)		USA (Day)	
	Train	Input	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
MeanPred	-	C	65.80	2.44	3.13	1.14	5.66	1.13	13.85	3.02
IDW	ACT	ACT	39.94	2.51	2.56	0.95	4.56	1.05	**10.24**	2.57
	AC	AC	351.73	2.85	2.66	0.95	7.33	1.61	23.21	5.29
	C	C	25.83	2.77	**2.31**	0.98	4.35	0.91	10.32	2.60
RF	ACT	ACT	**22.24**	2.81	2.37	0.95	4.18	0.68	10.73	2.89
	AC	AC	77.30	2.67	2.69	0.98	6.05	0.93	13.93	3.20
	C	C	22.25	2.77	2.34	0.89	4.12	0.68	10.82	2.85
XGBoost	ACT	ACT	33.24	2.87	2.55	0.95	4.62	1.01	11.51	3.05
	AC	AC	65.04	2.55	2.90	0.98	6.03	0.84	14.19	3.32
	C	C	29.73	2.76	2.71	1.05	**4.09**	0.67	11.66	3.16
NSGP	ACT	ACT	29.11	3.84	2.57	1.09	4.41	0.89	10.39	2.69
	ACT	C	194.96	1.63	13.02	0.72	14.68	0.63	*26.43	*3.08
	AC	AC	69.75	3.65	2.89	0.90	5.99	0.95	*12.65	*2.33
	AC	C	37.46	4.63	3.17	1.12	5.25	1.22	*21.02	*2.69
	C	C	170.99	9.31	12.74	0.55	13.51	0.72	27.81	3.67
Graphsage	AC	C	38.63	3.89	2.96	1.25	5.37	1.13	11.66	3.29
	C	C	38.68	4.12	3.13	1.24	5.68	1.46	12.75	4.06

Table 4 shows the complete benchmark for Spatio-temporal Missing data Imputation for different train and input configurations. Missing data imputation is briefly discussed in § 4.4 in the main paper. The benchmarks for NSGP algorithm for some configurations for USA dataset cannot be computed yet due to resource constraints. We will do this soon and update as applicable. As per our understanding, this information will not impact the analysis presented so far. The traditional and powerful RF (Random Forest) algorithm outperforms all other algorithms and methods.

Table 4: Missing Data Imputation RMSE for different configurations.

Algo	Config		Delhi (Day)		Canada (Day)		Canada (Year)		USA (Day)	
	Train	Input	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
MeanPred	-	C	65.80	2.44	3.13	1.14	5.66	1.13	13.85	3.02
IDW	ABCT	ABCT	40.06	2.51	2.56	0.95	4.56	1.05	10.19	2.57
	ABC	ABC	399.44	1.14	2.69	0.93	7.92	1.47	68.63	8.00
RF	ABCT	ABCT	**22.26**	2.85	**2.34**	0.93	**4.22**	0.67	**9.42**	2.60
	ABC	ABC	78.90	2.71	2.70	0.96	6.21	0.96	14.09	3.13
XGBoost	ABCT	ABCT	33.46	2.87	2.53	0.91	4.63	1.02	10.23	2.74
	ABC	ABC	67.66	2.55	2.94	0.96	6.19	0.87	13.84	3.12
NSGP	ABCT	ABCT	29.06	3.64	2.52	0.95	4.40	0.85	9.62	2.46
	ABC	ABC	71.27	3.16	2.81	0.91	6.09	0.88		
	ABC	C	171.94	8.08	12.71	0.53	13.29	0.94		
	ABCT	C	194.98	1.55	12.90	0.60	14.58	0.68		
	ABT	C	195.86	3.00	13.03	0.61	14.68	0.95		
	AB	C	37.63	3.87	4.15	0.92	5.43	1.09		
Graphsage	ABC	C	38.53	2.94	3.15	1.30	5.46	1.11	11.78	3.56
	AB	C	38.48	2.86	3.13	1.25	5.41	1.08	11.59	3.15

Table 5 shows the complete benchmark for Spatio-temporal Forecasting for different configurations. A subset of these benchmarks is presented in Fig. 6 and discussed in § 4.4 in the main paper.

Table 5: Forecasting RMSE for different configurations.

Algo	Config	Delhi (Day)	Canada (Day)	Canada (Year)	USA (Day)
IDW	ABT	86.52	5.65	8.31	14.61
	AB	270.73	**5.73**	11.23	69.20
RF	ABT	110.49	5.90	8.45	14.23
	AB	89.54	6.11	10.80	14.58
XGBoost	ABT	102.68	6.69	8.23	14.25
	AB	**84.15**	6.51	9.84	14.52
NSGP	ABT	95.83	5.76	**8.01**	**13.65**
	AB	86.34	6.08	10.22	14.34
ARIMA	ABT	148.86	13.87	12.85	20.12
nBeats	ABT	106.41	10.88	11.84	17.05

NSGP Variance

Non-stationary GP models provides us with uncertainty (variance) values around the expected mean PM2.5 value for each expected spatio-temporal location. We find that the average variance value for Delhi dataset is huge as compared to Canada (Day) experiments. It is more challenging for a model or algorithm to correctly understand and predict the PM values for Delhi dataset. Even the USA dataset with data over a big region does not exhibit such complexity for the algorithms.

Table 6: NSGP Variance.

	Delhi (Day)	Canada (Day)	Canada (Year)	USA (Day)
Spatio-temporal Interpolation	118.73	17.29	72.94	76.34
Missing Data Imputation	142.51	20.34	113.37	72.58
Forecasting	77.38	19.96	60.89	59.76

C Anova Tests Analysis for Low Cost Sensor

In continuation to the data quality analysis presented in § 3.2, we performed Anova Tests over the data collected by DustTrak and our Low Cost Mobile sensor devices at the same location. ANOVA Navidi [2009], Analysis of Variance, is a strong statistical factorial technique which involves one dependent variable known as response variable and one or more independent variables known as factors. The factors have different levels called treatments. The ANOVA tests compare two types of variation, the variation between the sample means and the variation within the samples.

Two-way ANOVA test between DustTrak reference sensor and our low-cost mobile sensor

In relation to our low cost sensor scenario, the observed $\mathrm{PM}_{2.5}$ values are dependent on the sensor *Type* (DustTrak vs Low Cost) and the time(*Day*) of observation. As we have two factors, we need to perform two-way ANOVA test. For the *Day* factor, we take the hourly $\mathrm{PM}_{2.5}$ mean samples grouped over each day (24 hours) of observations.

Two-way ANOVA tests three *null* hypotheses

(a) the means of observations grouped by factor *Type* are same

(b) the means of observations grouped by factor *Day* are same

(c) there is no interaction between the two factors *Type* and *Day*

Two-way ANOVA Assumptions

We make the standard assumptions of completeness, balanced design, normal distribution, similar variance, and sufficient replicates per treatment for validating ANOVA hypotheses. We take one device per sensor *Type* and same number (11) of *Day* as treatments under the two factors, with each *Type* and *Day* containing $\mathrm{PM}_{2.5}$ samples. Fig. 8 shows the box-plot diagram with similar standard deviation for the DustTrak and our Low cost mobile sensors.

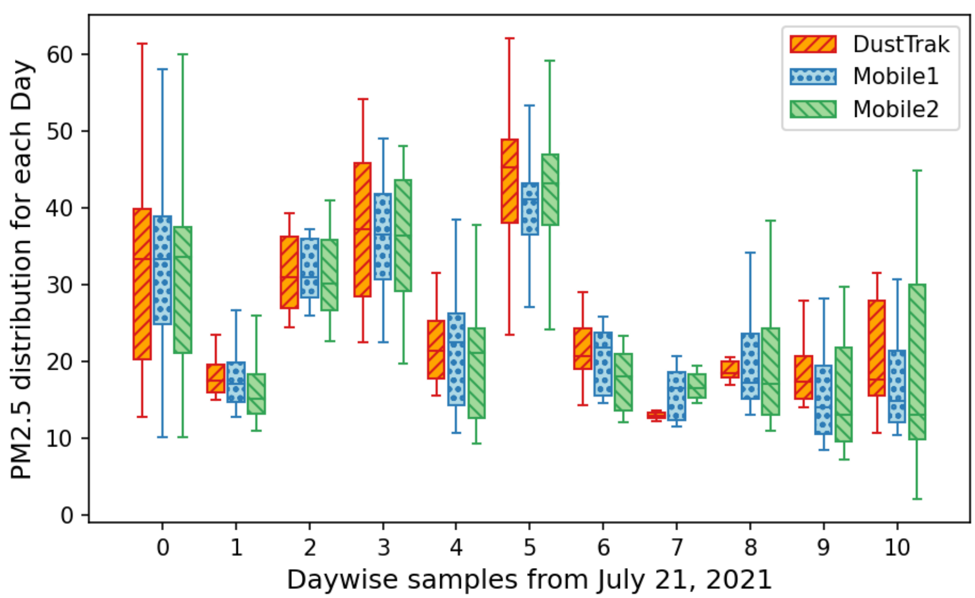


Figure 8: Mean and Standard Deviation for DustTrak and our Low Cost Mobile sensors.

Table 7: Two-way ANOVA test for DustTrak Reference Sensor vs Our Low Cost Sensor Mobile Sensor 1

Effect	Source	df	SumSq	MeanSq	F	p-value	Significance
Main	*Type*	1	197.84	197.84	2.36	0.1248	Holds hypo (a)
	Day	10	30204.98	3020.50	36.10	< 0.0001	Reject hypo (b)
Interaction	*Type*Day*	10	261.76	26.18	0.31	0.9778	Holds hypo (c)
Error	Residual	444	37147.11	83.66			

Interpreting two-way ANOVA results

Table 7 shows the two-way ANOVA test results for DustTrak and our Low Cost Mobile sensor. As per Seltman [2018], the *SumSq* column represents the sum of squared deviations for each *Source* of variation. Each *Source* has a *df* (degrees of freedom) which is a measure of the number of independent pieces of information present in the deviations that are used to compute the corresponding *SumSq*. Each *MeanSq* is a variance estimate and the *SumSq* divided by the *df* for that *Source*.

Each F-statistic is the ratio of two *MeanSq* values. For the main effects, *Type* and *Day*, the denominators are all MSE which are pure estimates of group variance, unaffected by the validity of the null hypothesis. Each F-statistic is compared against it's null sampling distribution to compute a *p-value*. Interpretation of each of the *p-values* depends on the corresponding null hypothesis.

In the presence of an interaction (*Type*Day*), the *p-value* for the interaction is most important and the main effects *Type* and *Day* p-values would be ignored if the interaction is significant. This is mainly because if the interaction is significant, then some changes in both explanatory variables (*Type* and *Day*) must have an effect on the outcome $PM_{2.5}$, regardless of the main effect *p-values*. The null hypothesis for the interaction F-statistic supports an additive relationship between the two explanatory variables, *Type* and *Day*, in their effects on the outcome $PM_{2.5}$. If the *p-value* for the interaction is less than α (usually 0.05), then we have a statistically significant interaction.

As we have a non-significant interaction $F_{1,10} = 0.31$ with *p-value* $= 0.9778$ which is greater than $\alpha = 0.05$, the null hypothesis (c) holds and the *p-values* for the main effects are valid for consideration. So, we can see that the *Day* has a significant *p-value* and thus it rejects the null hypothesis (b) meaning that there is impact of different *Day*'s observation on the observed $PM_{2.5}$ sample. This outcome aligns with a common understanding regarding the varying pollution across different days.

The analysis for the main effect sensor *Type* is more encouraging. It has a non-significant *p-value* $= 0.1248$ which holds the null hypothesis (a) that the means of the observations of the two device *Types*, DustTrak and our Low Cost Mobile sensor, are same. Hence, our Low Cost Mobile device can be effectively used to collect $PM_{2.5}$ observations in place of the expensive DustTrak sensors.

One-way ANOVA test between DustTrak reference sensor and our low-cost mobile sensor

Though the two-way ANOVA results hold for the main effects, we still perform one-way ANOVA test for the main effect *Type* (DustTrak vs Low Cost) for the observed $PM_{2.5}$ values. We ignore the *Day* factor in this analysis, so the $PM_{2.5}$ samples are only attributed with the *Type* factor. One-way ANOVA tests for the hypothesis (a) as of two-way ANOVA and with the standard assumptions of normal distribution and similar variance.

Table 8 presents the results for one-way ANOVA, which too shows *Type* factor to have a non-significant *p-value* $= 0.2445$ which holds the null hypothesis (a). Hence with similar means of the observations, our Low Cost Mobile device can replace the expensive DustTrak sensors.

Table 8: One-way ANOVA test for DustTrak Reference Sensor vs Our Low Cost Sensor Mobile Sensor 1

Effect	Source	df	SumSq	MeanSq	F	p-value	Significance
Main	*Type*	1	197.84	197.84	1.36	0.2445	Holds hypothesis (a)
Error	Residual	464	67613.85	145.72			

Two-way ANOVA test for our Low Cost device replaceability

We also show that our Low Cost Mobile devices are replaceable by each other. We perform two-way ANOVA tests between our Low Cost Mobile devices and the results are presented in Table 9.

Table 9: Two-way ANOVA test for Our Low Cost Sensor Mobile Sensor 1 vs 2

Effect	Source	df	SumSq	MeanSq	F	p-value	Significance
Main	*Type*	1	145.65	145.65	1.65	0.1991	Holds hypothesis (a)
	Day	10	31204.66	3120.47	35.43	< 0.0001	Reject hypothesis (b)
Interaction	*Type*Day*	10	148.46	14.85	0.17	0.9982	Holds hypothesis (c)
Error	Residual	450	39632.11	88.07			

As the *p-value* for the interaction is non-significant, main effects are valid. Likewise *Day* factor rejects hypothesis (b) and importantly *Type* factor holds hypothesis (a), allowing our Low Cost devices to replace each other as applicable.

References

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *31st NeurIPS Conference*, 2017.

William Navidi. *Statistics for Engineers and Scientists*. McGraw-Hill, 2009

Howard Seltman. *Experimental Design and Analysis.* Carnegie Mellon University, 2018.

D Letters of Approval / Certifications from authorities

D.1 ICAT EMC certification

ICAT EMC certification of our instrument verifying that it doesn't interfere with the bus's electro-mechanical properties.

ICAT
Innovation • Service • Excellence

INTERNATIONAL CENTRE FOR AUTOMOTIVE TECHNOLOGY

[A Division of NATRiP Implementation Society (NATIS), Govt. of India]

Non-Transferable

TEST REPORT

ULR No. : T C 5 3 6 0 1 9 0 5 0 0 0 0 3 9 6 F **Date: 12-09-2019**

Test Report No.: C D 0 M 0 0 4 7 4

1.0 NAME AND ADDRESS OF THE CUSTOMER **: Computer Science Department IIT Delhi**
Room 514, Bharti Building IIT Delhi Hauz Khas New Delhi-110016
New Delhi INDIA

CUSTOMER REFERENCE **:CCDCSDIOHEMC70648; Dated: 17-07-2019**

2.0 DESCRIPTION OF TEST PROPERTY :

DUT Name : Aerogram
Model Name : Aerogram
Voltage system: 24V DC
Part No. : 01
Drawing No : EzioMotiv – V1.0

3.0 DATE OF RECEIPT OF TEST PROPERTY: 03-09-2019

4.0 CONDITION OF SAMPLE: Sample received in good condition.

5.0 TEST OBJECTIVE: To conduct tests as mentioned in Sr. No. 13 on DUT as per AIS004 - Part 3 as amended up to April 2015.

6.0 TEST METHOD: As per AIS004 - Part 3 as amended up to April 2015.

7.0 ANY DEVIATION OR EXCLUSION FROM TEST METHOD: NA

8.0 FUNCTIONAL VERIFICATION: DUT was powered on with 27 V DC (24 V DC system)for all tests as per AIS 004-3 as amended up to April 2015. Performance was observed by monitoring data strings before and after each test and during the test, and power LED of the device was monitored.

9.0 CONCLUSION: The component mentioned in Sr. No. 2 above meets the requirements as per AIS004 - Part 3 as amended up to April 2015.

10.0 TEST DESCRIPTION: EMC/EMI testing as per AIS004 - Part 3 as amended up to April 2015 on the DUT.

11.0 DATE OF PERFORMANCE OF TEST: 09-09-2019 - 10-09-2019

12.0 LOCATION OF TEST: ICAT EMC LAB, Manesar

13.0 TEST RESULTS:

Sr. No.	TEST TITLE	SAMPLE ID	REF. STD.	OBSERVATIONS/RESULTS
1	Radiated Emission	ICAT/EMC/70648/01	AIS004- Part 3: 2009 as amended up to April 2015	*Refer Annexure-I*
2	Radiated Immunity			*Refer Annexure-II*
3	Conducted Transient Emission			*Refer Annexure-III*
4	Conducted Transient Immunity			*Refer Annexure-IV*

Refer Annexure-V for test setup photographs.

Prepared By	Checked By		Approved By	
JEEVAN PAL Deputy Manager	NIKHIL GROVER Manager	INTERNATIONAL CENTRE FOR AUTOMOTIVE TECHNOLOGY ICAT MANESAR	S. K. KALIA CBO-ICAT Centre-II	M00474 Page 1 of 6 + Dwg.-01 [70648]

Office Address : Centre-I : Plot No.-26, Sector-3, HSIIDC, IMT-Manesar, Gurugram-122050. Haryana (India)
Centre-II : Plot No.-01, Sector-M-11, HSIIDC, IMT-Manesar, Gurugram-122050. Haryana (India)
Phone : 0124-4586111, Fax : +91-124-2290005. E-mail: team@icat.in, Website : www.icat.in
(An ISO 9001, ISO 14001 and ISO 45001 certified, scope wise NABL accredited and BIS recognised Test House)

T	C	5	3	6	0	1	9	0	5	0	0	0	0	3	9	6	F	Date: 12-09-2019
C	D	0	M	O	0	4	7	4										

ICAT
Innovation • Service • Excellence

14.0 CLASSIFICATION OF FUNCTIONAL STATUS:

CLASSIFICATION OF FUNCTIONAL STATUS AS PER (A.4) ANNEX-A, ISO 7637-2:2004:	
CLASSES	**DESCRIPTION**
CLASS A	All functions of the device/system perform as designed during and after the test.
CLASS B	All functions of the device/system perform as designed during the test. However, one or more may go beyond the specified tolerance. All functions return automatically to within normal limits after the exposure is removed.
CLASS C	One or more functions of a device/system do not perform as designed during the test but return automatically to normal operation after the exposure is removed.
CLASS D	One or more functions of a device/system do not perform as designed during the exposure and do not return to normal operation until exposure is removed and the device/system is reset by simple 'operator/use' action.
CLASS E	One or more functions of a device/system do not perform as designed during and after exposure and cannot be returned to proper operation without repairing or replacing the device/system.

15.0 LIST OF EQUIPMENTS USED IN THE TEST AND CALIBRATION DETAILS:

Lab ID	Name of Instruments	Manufacturer	Model (S. No.)	Calib. due date
		Radiated Emission		
ICAT/EMC/TR - 01	EMI Test Receiver	Rohde and Schwarz	ESU-8 (100290)	03/05/2020
ICAT/EMC/EPA-01	External Preamplifiers	TDK RF Solutions	PA-02-001-100 (121054)	03/05/2022
ICAT/EMC/OFBA-04	Biconical Antenna with polarization adaptor	TDK RF Solutions	PBA2030 (130818)	09/05/2021
ICAT/EMC/OFBA-05	Broadband Log periodic Antenna	TDK RF Solutions	PLP 3003 (130830)	09/05/2021
ICAT/EMC/AN-01	LISN	Schwarzbeck	NNBM8124 (8124-649)	03/05/2022
ICAT/EMC/AN-02		Schwarzbeck	NNBM8124 (8124-650)	03/05/2022
		Radiated Immunity		
ICAT/EMC/SG-01	Signal Generator	Agilent Technologies	N5183A-520 (50140523)	29/04/2022
ICAT/EMC/SG-03		Agilent Technologies	SMB100A (103955)	30/04/2022
ICAT/EMC/AN-01	LISN	Schwarzbeck	NNBM8124 (8124-649)	03/05/2022
ICAT/EMC/AN-02		Schwarzbeck	NNBM8124 (8124-650)	03/05/2022
ICAT/EMC/CIP-02	Current injection probe	FCC	F -140 (130055)	-
ICAT/EMC/OFBA-07	V Log Array Antenna	TDK RF Solutions	VLA-8001 (130835)	-
ICAT/EMC/OFBA-10	Horn Antenna	TDK RF Solutions	ATH800M5GA (0337348)	-
ICAT/EMC/PM-01	RF Power Meter	Agilent Technologies	N1914A (MY50000499)	30/04/2022
ICAT/EMC/PM-03		Agilent Technologies	N1912A (MY54010017)	30/04/2022
ICAT/EMC/AMP-01	Amplifier	AR	500W1000A (0335094)	-
ICAT/EMC/AMP-02		AR	2500A225, Sr. No. 338773	-
ICAT/EMC/AMP-03		AR	500T1G2 (0336388)	-
ICAT/EMC/APS-01	Average Power sensor (9kHz-6GHz)	Agilent Technologies	E9304 (S.No. MY51020021)	30/04/2020
ICAT/EMC/APS-02		Agilent Technologies	E9304 (S.No. MY51030004)	30/04/2020
ICAT/EMC/PS-01	Power Sensor (50MHz-18GHz)	Agilent Technologies	N1921A (MY53380017)	30/04/2020
ICAT/EMC/PS-02		Agilent Technologies	N1921A (MY53380020)	30/04/2020
ICAT/EMC/FP-06	Field Probe	AR	FL7018 (0334718)	30/09/2020
		Conducted Transient Emission		
ICAT/EMC/PG/05	Voltage drop simulator	EM test	VDS 200N100	21/01/2020
ICAT/EMC/DSO/01	Digital Storage Oscilloscope	EM test	DSO9254A	21/01/2020
ICAT/EMC/AN/01	Single line artificial network	EM test	AN 200N100	21/01/2020
ICAT/EMC/SW/01	Electronic switch	EM test	BS 200N100	21/01/2020
ICAT/EMC/MR/01	Matching resistor for transient generators	EM test	CAISO	21/01/2020
		Conducted Transient Immunity		
ICAT/EMC/PG/02	Ultra-Compact Simulator	EM test	UCS 200N100	21/01/2020
ICAT/EMC/PG/05	Voltage drop simulator	EM test	VDS 200N100	21/01/2020
ICAT/EMC/PG/05	Load dump simulator	EM test	LD200N	21/01/2020

Prepared By		Checked By	
Jeevan	INTERNATIONAL CENTRE FOR AUTOMOTIVE TECHNOLOGY ICAT Innovation • Service • Excellence MANESAR	Nhovey	Page 2 of 6 + Dwg.-01 [70648]
JEEVAN PAL		NIKHIL GROVER	
Deputy Manager		Manager	

T	C	5	3	6	0	1	9	0	5	0	0	0	0	3	9	6	F	Date: 12-09-2019

C	D	0	M	O	0	4	7	4

Innovation • Service • Excellence

Annexure – I

1.0 Measurement of Radiated Emissions:

1.1 Test Condition:

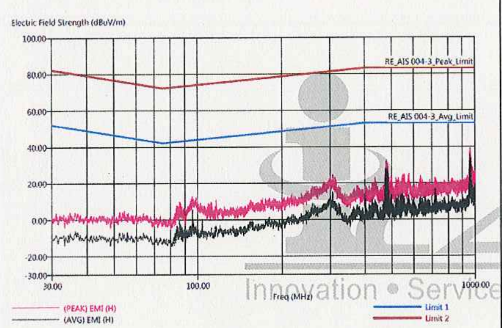
Operating Condition	Powered ON

1.2 Test Specifications:

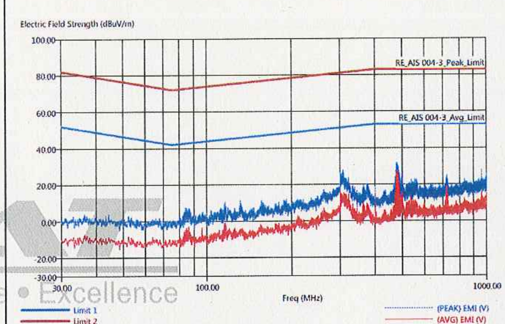
Frequency Range	30MHz - 1000MHz
Step Size	50kHz
Bandwidth	120kHz
Measurement time	5ms
Antenna	30MHz - 300MHz: Bi-conical antenna, 300MHz - 1000MHz: Log-Periodic antenna
Antenna Polarization	Horizontal and Vertical
Antenna Location	In front of centre of harness
Antenna Distance	1 meter
Detector	Peak and Average
Harness length	1700mm

1.3 Test Graphs:

Graph for Horizontal Data



Graph for Vertical Data



1.4 Test Requirements:

Radiated Emissions measured should be within limits defined in AIS004- Part 3: 2009 as amended up to April 2015.

1.5 Test Observations /Results:

Radiated Emissions measured are within limits.

Prepared By		Checked By	
Jeevan	INTERNATIONAL CENTRE FOR AUTOMOTIVE TECHNOLOGY ★ MANESAR ★ ICAT Innovation • Service • Excellence	Grover	Page 3 of 6 + Dwg.-01 [70648]
JEEVAN PAL		NIKHIL GROVER	
Deputy Manager		Manager	

T	C	5	3	6	0	1	9	0	5	0	0	0	3	9	6	F	Date: 12-09-2019
C	D	0	M	O	0	4	7	4									

Innovation • Service • Excellence

Annexure – II

2.0 Radiated Immunity Test:

2.1 Bulk Current Injection (BCI):

2.1.1 Test Condition:

Operating Mode	Powered ON

2.1.2 Test Specifications:

Frequency Range	20MHz – 80MHz
Step Size	5%
Current Severity Level	60mA
Dwell Time	2s
Harness length	1700mm
Current probe position	150mm from DUT
Test Method	Substitution (open loop)
Modulation	Amplitude modulation with 1 kHz modulating frequency and 80 % modulation depth.

2.1.3 Test Observations /Results:

S. No.	Frequency Range	Modulation	Acceptance Criteria	Observation/Result
1.	20MHz to 80MHz	Amplitude Modulation (AM)	No deviation in performance of DUT should be observed during test	No deviation observed

2.2 Absorber Lined Shielded Enclosure (ALSE) method:

2.2.1 Test Condition:

Operating Mode	Powered ON

2.2.2 Test Specifications:

Frequency Range	80MHz – 2000MHz
Step Size	80-400MHz: 5%, 400-2000MHz: 2%
Field Severity Level	30V/m
Dwell Time	2s
Harness length	1700mm
Antenna	80MHz – 1000MHz: V log array antenna, 1000MHz – 2000MHz: Horn antenna
Antenna Polarization	Vertical
Antenna Location	80MHz – 1000MHz: in front of centre of harness, 1000MHz – 2000MHz: in front of DUT
Antenna Distance	1 meter
Test Method	Substitution
Modulation	80MHz – 800MHz: Amplitude modulation with 1 kHz modulating frequency and 80 % modulation depth 800MHz – 2000MHz: Pulse modulation: Ton: 577µs, period: 4600µs

2.2.3 Test Observations /Results:

Sr. No.	Frequency Range	Modulation	Antenna Polarization	Acceptance Criteria	Observation/ Result
1.	80MHz to 800MHz	Amplitude Modulation	Vertical	No deviation in performance of DUT should be observed	No deviation observed
2.	800MHz to 1000MHz	Pulse Modulation			
3.	1000MHz to 2000MHz	Pulse Modulation			

Prepared By		Checked By	
JEEVAN PAL Deputy Manager	INTERNATIONAL CENTRE FOR AUTOMOTIVE TECHNOLOGY ICAT Innovation • Service • Excellence MANESAR	NIKHIL GROVER Manager	Page 4 of 6 + Dwg.-01 [70648]

T	C	5	3	6	0	1	9	0	5	0	0	0	0	3	9	6	F
C	D	0	M	O	0	4	7	4									

Date: 12-09-2019

Innovation • Service • Excellence

Annexure – III

3.0 Measurement of Conducted Transient Emissions:

3.1 Test Condition:

Operating Condition	Powered ON

3.2 Test Observations/Results:
24V System

Sr. No.	Supply Polarity	Limits as per AIS 004:Part 3	Observation	Results
Fast transient				
1.	DUT ON to OFF	Positive: +150V Negative: -400V	Positive Transient: No Significant Transient Negative Transient: No SignificantTransient	Within Limits
2.	DUT OFF to ON		Positive Transient: 42.0 V NegativeTransient: No SignificantTransient	Within Limits
Slow transient				
3.	DUT ON to OFF	Positive: +150V Negative: -400V	Positive Transient: No Significant Transient Negative Transient: No Significant Transient	Within Limits
4.	DUT OFF to ON		Positive Transient:43.32V NegativeTransient: No SignificantTransient	Within Limits

Annexure – IV

4.0 Immunity to Transient Disturbances Conducted along Supply Lines as per AIS 004-3 as amended up to April 2015 following ISO 7637-2:2004:

4.1 DUT Condition:

Operating Condition	Powered ON

4.2 Test Requirements and Observations/Results:
24V System:

Test Pulse	Severity Level	Acceptance Criteria	Achieved Class	Observations	Results
Pulse 1	III	Class C	Class C	Reset Observed during pulse injection	Satisfactory
Pulse 2a		Class B	Class A	No deviation in performance observed	Satisfactory
Pulse 2b		Class C	Class C	Reset Observed during pulse injection	Satisfactory
Pulse 3a		Class A	Class A	No deviation in performance observed	Satisfactory
Pulse 3b		Class A	Class A	No deviation in performance observed	Satisfactory
Pulse 4		Class C	Class C	Reset Observed during pulse injection	Satisfactory

Prepared By		Checked By	
Jeevan	INTERNATIONAL CENTRE FOR AUTOMOTIVE TECHNOLOGY ICAT Innovation • Service • Excellence ★ MANESAR ★	Nhgrover	Page 5 of 6 + Dwg.-01 [70648]
JEEVAN PAL		NIKHIL GROVER	
Deputy Manager		Manager	

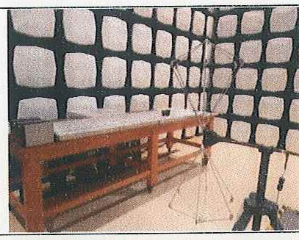
T	C	5	3	6	0	1	9	0	5	0	0	0	0	3	9	6	F	Date: 12-09-2019
C	D	0	M	O	0	4	7	4										

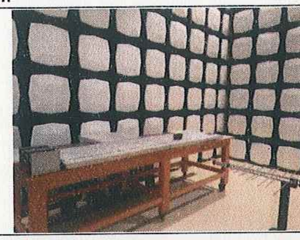
ICAT
Innovation • Service • Excellence

Annexure – V

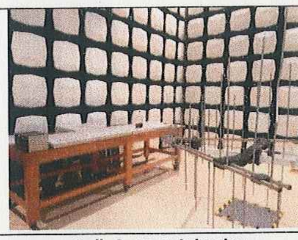
5.0 Test Setup and Test Circuitry:

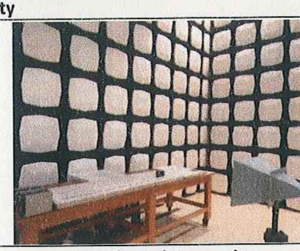
Radiated Emission



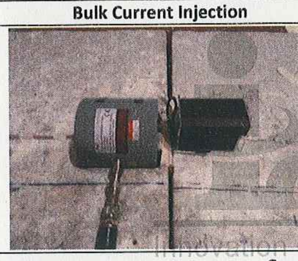


Radiated Immunity





Bulk Current Injection



Conducted Immunity

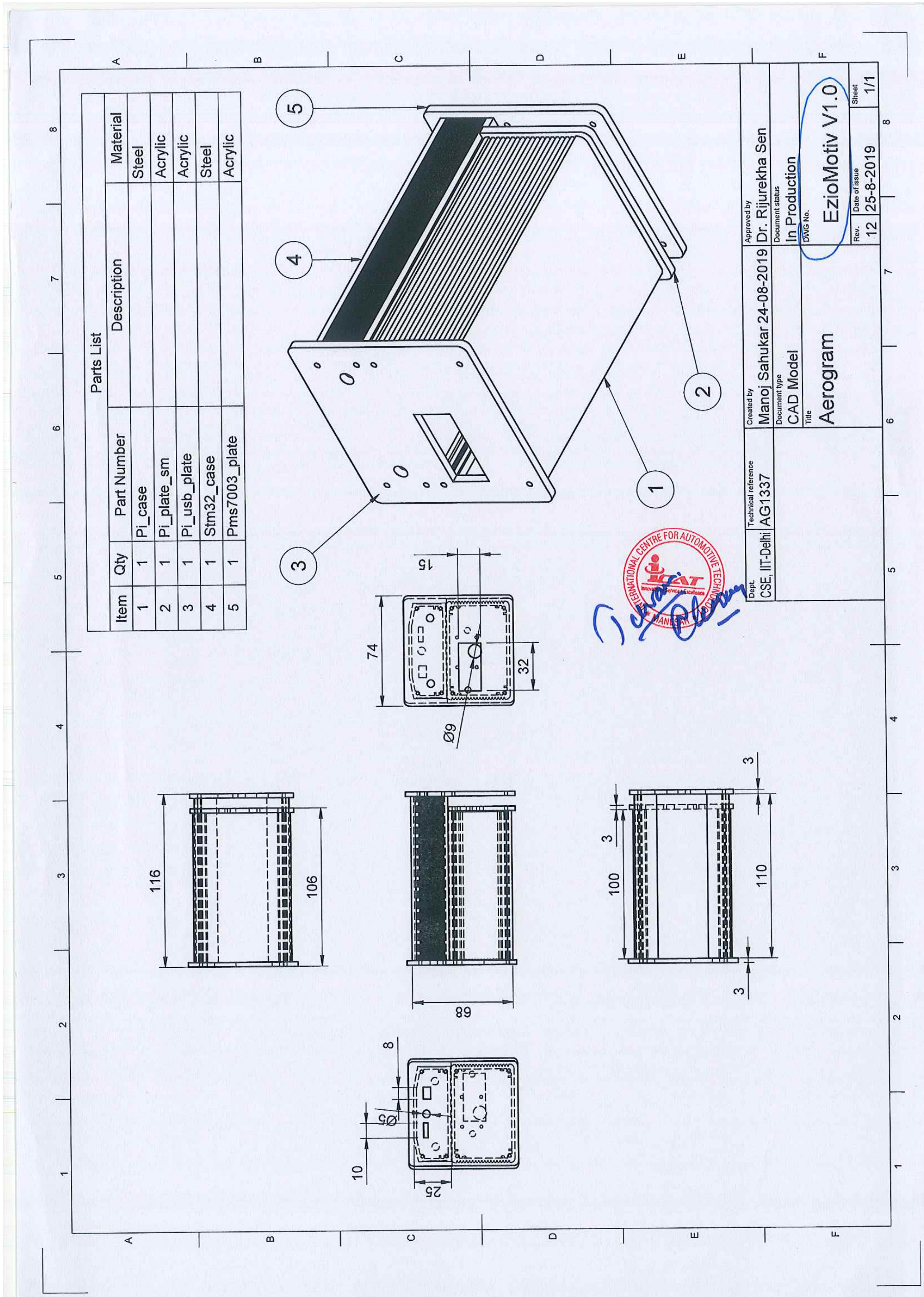
Conducted Transient Emission

Setup for Fast Transient

Setup for Slow Transient

---------- END OF REPORT ----------

Prepared By		Checked By	
JEEVAN PAL Deputy Manager	INTERNATIONAL CENTRE FOR AUTOMOTIVE TECHNOLOGY ICAT MANESAR	NIKHIL GROVER Manager	Page 6 of 6 + Dwg.-01 [70648]


Parts List
Item | Qty | Part Number | Description | Material
1 | 1 | Pi_case | | Steel
2 | 1 | Pi_plate_sm | | Acrylic
3 | 1 | Pi_usb_plate | | Acrylic
4 | 1 | Stm32_case | | Steel
5 | 1 | Pms7003_plate | | Acrylic
Dept.
CSE, IIT-Delhi
Technical reference
AG1337
Created by
Manoj Sahukar 24-08-2019
Approved by
Dr. Rijurekha Sen
Document type
CAD Model
Document status
In Production
Title
Aerogram
DWG No.
EzioMotiv V1.0
Rev.
12
Date of issue
25-8-2019
Sheet
1/1

D.2 Delhi Integrated Multi-Modal Transit System (DIMTS) letter of support

DIMTS

We help people move

Ref: DIMTS/TP/2018/2756 Dated: June 21, 2018

To,
Department of Science & Technology
Delhi -

Subject : Letter of Support for the Proposed Research study.

On behalf of DIMTS, we will extend our support to Profs Pravesh Biyani and Rijurekha Sen for their research proposal related to "Vehicle mounted Particulate Matter (PM) sensing in Delhi-NCR".

DIMTS runs more than 1600 non air-conditioned cluster buses on various routes in the Delhi region. We will facilitate the use of some of the vehicle fleet as needed by the researchers for pilot studies as they build and test their vehicle mounted sensing system.

Pollution being a pressing problem in Delhi-NCR, partnering with this research effort in a meaningful way will be very exciting for DIMTS.

Thanking you.

Yours faithfully,

Samir Sharma

Samir Sharma
Vice President - Transport Planning

DELHI INTEGRATED MULTI-MODAL TRANSIT SYSTEM LTD.
(A joint venture of the Govt. of NCT of Delhi and IDFC Foundation)
An ISO 9001:2015, ISO 14001:2015, OHSAS 18001:2007 & ISO 27001 certified and CMMI appraised company
CIN No. U60232DL2006PLC148406
REGD. OFF.: 1ST FLOOR, MAHARANA PRATAP ISBT BUILDING, KASHMERE GATE, DELHI 110 006 (INDIA)
Tel: +91 11 43090100 • Fax: +91 11 23860966 • Email : info@dimts.in • Web: www.dimts.in

D.3 Delhi Pollution Control Committee (DPCC) letter of Support

Delhi Pollution Control Committee

5th Floor I.S.B.T. Building Complex Kashmere Gate Delhi 110006

Visit us at :http: //dpcc.delhigovt.nic.in

F. No. DPCC/(12)(1)(260)Lab(A)2020/2203 Date: 27/1/2020

To,

Dr. Rijurekha Sen
Department of Computer Science,
IIT Delhi, Hauz Khaz,
New Delhi-110016

Subject- Support Letter for Vehicle Mounted Low Cost PM Monitoring in Delhi

Madam,

With reference to your E-mail and telephonic discussion this organization is interested to know feasibility of Vehicle Mounted Low Cost PM Monitoring in Delhi and willing to share data generated by DPCC Ambient Air Quality Network to assess error percentage of Low Cost System.

(Dr. M. P. George)
Scientist

Dr. M. P. GEORGE
Scientist

D.4 Delhi Ministry of Transport (MOT) Permission

GOVERNMENT OF NCT OF DELHI
TRANSPORT DEPARTMENT
(CLUSTER & DTC SECRETARIAT)
5/9, UNDER HILL ROAD, DELHI – 110 054

No. F.10/STA/Policy /Tpt./ 2011/ 333/40631 **Date:** 17/08/2020

To

The CEO,
Delhi Integrated Multi Modal Transit System Ltd.,
8^{th} Floor, Block-1, Delhi Technology Park,
Shastri Park, Delhi-110053.

Subject: Request for permission to install pollution sensing units in Cluster buses as a part of R&D Project by IIT, Delhi.

Sir,

Kindly refer to your letter no. DIMTS/Road Transport/2019/4398, dated 05.11.2019, on the abovementioned subject. DIMTS had requested for a formal approval to install pollution sensing units in 10 Cluster buses by CSE IIT, Delhi.

In this context, I am directed to convey the approval of Hon'ble Minister (Transport) for installing of pollution sensing units in 10 Cluster buses of the Kushak Nalah Depot by CSE, IIT Delhi.

Yours faithfully,

(Subodh Kumar)
Deputy Commissioner
(Cluster & DTC Sectt.)

Copy to:
1. Dy. Commissioner (PCD) with reference to U.O. NO. 23(1471)/CAP/TPT/PCD/ 2018/ 1595/87542 dated 26.11.2019.
2. Ms. Rijurekha Sen, Assistant Professor, CSE, IIT, Delhi.
3. M/s. Indraprastha Logistics Pvt. Ltd, 80/2, Ground Floor Govindpuri Kalkaji New Delhi-110019

D.5 Letter of funding: SCIENCE & ENGINEERING RESEARCH BOARD (SERB), INDIA

FILE NO. IMP/2018/001481

SCIENCE & ENGINEERING RESEARCH BOARD (SERB)

(A statutory body of the Department of Science & Technology, Government of India)

5 & 5A, Lower Ground Floor
Vasant Square Mall
Plot No. A, Community Centre
Sector-B, Pocket-5, Vasant Kunj
New Delhi-110070

Dated: 29-Mar-2019

ORDER

Domain: **Information & Comm. Technology**
Subject: Financial Sanction of the research project titled "**Scalable Spatio-Temporal Measurement and Analysis of Air Pollution Data for Delhi-NCR using Vehicle-Mounted Sensors** " under the guidance of **Dr. Rijurekha Sen, Department of** Computer Science, Indian Institute of Technology Delhi , Hauz Khas, New Delhi, DELHI-110016 and by **Dr. Pravesh Biyani,** Assistant Professor, Ece Dept, Indraprastha Institute Of Information Technology and by **Dr. Arnab Bhattacharya,** Associate Professor, Department Of Computer Science And Engineering, Indian Institute Of Technology Kanpur and by Dr. Sayan Ranu, Assistant Professor, Computer Science And Engineering, Indian Institute Of Technology Delhi - Release of 1st grant.

Sanction of **Science and Engineering Research Board (SERB)** is hereby accorded to the above mentioned project at a total cost of **Rs. 12746800/- (Rs.** One Crore Twenty Seven Lakh Forty Six Thousand Eight Hundred **Only)** with break-up of **Rs.** ***5500000*****/- under Capital (Non-recurring) head** and **Rs.*****7246800*****/- under General (Recurring) head** for a duration of 36 months. The items of expenditure for which the total allocation of **Rs. 12746800/-** has been approved are given below:

S. No	Head	Total (in Rs.)
A	**Non-recurring**	
1	Equipment -> Laptop -> Server -> Sensors	5500000
A'	*Total (Non-Recurring)*	***5500000***
B	**Recurring Items**	
1	Recurring - I : (Manpower) Recurring - II : (Consumables, Travel, Contingencies) Recurring - III : Scientific Social Responsibility	3888000 2200000 0
2	Recurring - IV : (Overhead Charges)	1158800
B'	*Total (Recurring)*	***7246800***
C	**Total cost of the project (A' + B')**	**12746800**

2. Sanction of the **SERB** is also accorded to the payment of **Rs. 5500000/-** (Rupees Fifty Five Lakh only) under 'Grants for creation of capital assets' and **Rs. 2415000/-** (Rupees Twenty Four Lakh Fifteen Thousand only) under 'Grants-in-aid General' to **IRD, Indian Institute Of Technology Delhi, Hauz Khas, New Delhi** being the first installment of the grant for the year 2018-2019 for implementation of the said research project.

3. The expenditure involved is debitable to **Fund for Science & Engineering Research (FSER)**
This release is being made under Impacting Research Innovation and Technology (IMPRINT-2). (PAC Information & Communication Technology)

4. The Sanction has been issued to Indian Institute Of Technology Delhi, Hauz Khas, New Delhi with the approval of the competent authority under delegated powers on **28 March, 2019** and vide Diary No. **SERB/F/13078/2018-2019** dated **28 March, 2019**

5. Sanction of the grant is subject to the conditions as detailed in Terms & Conditions available at website (www.serb.gov.in).

6. Overhead expenses are meant for the host Institute towards the cost for providing infrastructural facilities and general administrative support etc. including benefits to the staff employed in the project.

7. While providing operational flexibility among various subheads under head Recurring-II, It should be ensured that not more than Rs. **450000** under Travel and Rs. **450000** under Contingency should be spent.

8. As per rule 211 of GFR, the accounts of project shall be open to inspection by sanctioning authority/audit whenever the institute is called upon to do so.

9. The sanctioned equipment would be procured as per GFR and its disposal of the same would be done with prior approval of SERB.

10. The release amount of **Rs. 7915000/-** (Rupees Seventy Nine Lakh Fifteen Thousand only) will be drawn by the Under Secretary of the SERB and will be disbursed by means of RTGS transaction as per their Bank details given below:

Account Name	**IRD ACCOUNTS IITD**
Account Number	10773572600
Bank Name & Branch	STATE BANK OF INDIA IIT BRANCH, IIT HAUZ KHAS, NEW DELHI - 110016
IFSC/RTGS Code	SBIN0001077
Email id of A/C Holder	arird@admin.iitd.ac.in
Email id of PI	riju@cse.iitd.ac.in

11.The institute will furnish to the SERB separate Utilization certificate(UCs) financial year wise to the SERB for Recurring (Grants-in-aid General) & Non-Recurring (Grants for creation of capital assets) and an audited statement of accounts pertaining to the grant immediately after the end of each financial year.

12. The institute will maintain separate audited accounts for the project. A part or whole of the grant must be kept in an interest earning bank account which is to be reported to SERB. The interest thus earned will be treated as credit to the institute to be adjusted towards further installment of the grant.

13. The project File no. IMP/2018/001481 should be mentioned in all research communications arising from the above project with due acknowledgement of SERB.

14. The manpower sanctioned in the project, if any is co-terminus with the duration of the project and SERB will have no liability to meet the fellowship and salary of supporting staff if any. beyond the duration of the project

15. As this is the first grant being released for the project, no previous U/C is required.

16. The institute may refund any unspent balance to SERB by means of a Demand Draft favoring "FUND FOR SCIENCE AND ENGINEERING RESEARCH" payable at New Delhi.

17. The organization/institute/university should ensure that the technical support/financial assistance provided to them by the Science & Engineering Research Board should invariably be highlighted/ acknowledged in their media releases as well as in bold letters in the opening paragraphs of their Annual Report.

18. In addition, the investigator/host institute must also acknowledge the support provided to them in all publications, patents and any other output emanating out of the project/program funded by the Science & Engineering Research Board.

Monika Agarwal

(Dr. Monika Agarwal)
Scientist E
ms.imprint@gmail.com

To,
Under Secretary
SERB, New Delhi
Copy forwarded for information and necessary action to: -

1.	The Principal Director of Audit, A.G.C.R.Building, IIIrd Floor I.P. Estate, Delhi-110002
2.	Sanction Folder, SERB , New Delhi.
3.	File Copy
4.	**Dr. Rijurekha Sen** **Department of Computer Science** **Indian Institute of Technology Delhi , Hauz Khas, New Delhi, DELHI-110016** **Email: riju@cse.iitd.ac.in** **Mobile: 919810591052** Dr. Pravesh Biyani **Ece Dept** **Indraprastha Institute Of Information Technology** **Dr. Arnab Bhattacharya** **Department Of Computer Science And Engineering** **Indian Institute Of Technology Kanpur** Dr. Sayan Ranu **Computer Science And Engineering** **Indian Institute Of Technology Delhi** (Start date of the project may be intimated by name to the undersigned. For guidance, terms & Conditions etc. Please visit www.serb.gov.in.)
5.	**IRD,** **Indian Institute Of Technology Delhi, Hauz Khas, New Delhi** (Receipt of Grant may be intimated by name to the undersigned)
6.	**Secretary,** **Department of Science & Technology** Ministry of Science and Technology Technology Bhavan, New Mehrauli Road, New Delhi-110016 Email: dstsec@nic.in
7.	**Secretary (Higher Education)** Ministry of Human Resource Development Shastri Bhavan, New Delhi- 110 001 Email: secy.dhe@nic.in

Monika Agarwal

(Dr. Monika Agarwal)
Scientist E
ms.imprint@gmail.com