Air Pollution Monitoring in Urban Areas Using Low Cost IoT Devices: ML based Calibration and Mobile Sensing

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Electronics and Communication Engineering
by Research

by

Spanddhana Sara
2020702021
spanddhana.sara@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
February 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Air Pollution Monitoring in Urban Areas using Low Cost IoT Devices: ML based Calibration and Mobile Sensing" by Spanddhana Sara, has been carried out under our supervision and is not submitted elsewhere for a degree.

Date	Advisor: Dr. Sachin Chaudhari

To

My Family and Friends

Acknowledgments

I remember being apprehensive and excited the day I started my internship at IIITH for my bachelor's thesis. I knew I had finally gone on a life-changing path of studying and investigating cutting-edge technologies. I will always be indebted to my adviser, Dr. Sachin Chaudhari, for giving me the opportunity to pursue my master's degree at this esteemed university and mentoring me the entire time. His vast knowledge and experience have been extremely helpful in defining my research and assisting me in achieving my academic objectives. I am incredibly appreciative of the numerous hours Dr. Chaudhari has dedicated to examining my work, offering criticism, and advising me while I conducted my study. His patience, knowledge, and encouragement were invaluable to my achievement, and I will be always thankful for his guidance.

I would like to thank Dr. Andrew Rebeiro-Hargrave, Dr. Samu Varjonen, Dr. Pak Lun Fung researchers from the University of Helsinki and Prof. Sasu Tarkoma, for their invaluable support and research collaboration. Their contributions have been instrumental in the success of our project. Throughout our collaboration, Andrew demonstrated exceptional expertise and dedication to his work, which greatly enriched our research. His insightful comments and suggestions helped us to refine our ideas and approaches, leading to significant improvements in our work. I am truly grateful for his contributions and look forward to future opportunities to work together.

I am grateful to acknowledge the Smart City Research Center and IHub data fellowship at IIITH; they have been instrumental in enabling me to pursue my research goals and has provided me with the financial support necessary to carry out my research.

I would also like to thank my project partners, Ishan Patwardhan, Ayu Parmar, Rajashekar Reddy, Ayush Kumar Dwivedi, Shreyash Gujar, and Om Kathalkar for their help and support. I also want to thank my friends Sai Usha, Jayati Narang, Sasanka GRS, Savitha Viswanath Kandala and others. I also want to express my gratitude to my lab mates for fostering an enjoyable and productive environment.

Last but not least, I want to express my gratitude to my family and friends for their love and support during this academic adventure. Their unfailing support and confidence in me have given me courage and motivation.

Abstract

The combination of Internet of Things (IoT) and Machine Learning (ML) technologies has the potential to completely transform air pollution monitoring and management. Low-cost air pollution measurement sensors have grown in popularity in recent years because they enable a low-cost solution to measure air quality in real time. However, these sensors frequently have accuracy concerns and must be calibrated to ensure that their results are valid. This thesis mainly focuses on calibrating low-cost sensors using ML models and the detection of pollutant hot spots in urban areas using IoT devices on a mobile platform.

In mobile monitoring of air pollution, IoT-enabled sensors assembled onto automobiles, and portable devices offer real-time data collecting when traveling between areas. This dynamic technique provides useful insights into real-world pollution fluctuations, allowing the identification of pollution sources and trends along traffic routes. The study highlights the importance of calibrating the IoT devices in mobile setting, even when the devices are calibrated in laboratory settings. Real-time ML algorithms can be used to calibrate sensor data based on location, weather, and additional relevant data.

In this thesis, a methodology is proposed for detecting emission spikes of $PM_{2.5}$, CO, and NO_2 in polluted urban environments employing portable low-cost sensors. Identification of harmful pollutant concentrations is achieved using two different IoT device types (MegaSense One and Prana Air) mounted on a mobile platform. Reliable identification of the $PM_{2.5}$, CO, and NO_2 emission spikes can be attained by driving through the city on different days. IoT device measurement errors were corrected by ML based calibration against a reference instrument co-located on a mobile platform. ML regression models like simple linear regression (LR), multivariate linear regression (MLR), polynomial regression (PR), support vector regression (SVR), decision tree regression (DT), random forest regression (RF) were applied to calibrate the devices.Among these models RF was the most suitable technique to reduce the variability between the IoT devices due to heterogeneity in the mobile sensing datasets. The spatial variability of $PM_{2.5}$, CO, and NO_2 harmful emission spikes at a resolution of $50\,\mathrm{m}$ were identified, but their intensity changes on a daily basis according to meteorological conditions. The data from the $PM_{2.5}$, CO, and NO_2 emission spikes at points of interests that disturb traffic flows clearly show the need for public education about when it is hazardous for persons with respiratory conditions to be outside, as well as when it is unsafe for young children and the elderly to be outside for extended periods of time. This detection strategy is adaptable to any mobile platform used by individuals traveling by foot, bicycle or drones in any metropolis.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
API	Application Program Interface
AQM	Air Quality Monitor
BLE	Bluetooth low energy
COVID-19	Corona Virus Disease - 2019
CPCB	Central Pollution Control Board
CPS	Cyber Physical Systems
DL	Deep Learning
ICT	Information and Communications Technology
IEEE	Institute of Electrical and Electronics Engineer
IIIT-H	International Institute of Information Technology - Hyderabad
IIoT	Industrial Internet of Things
IoT	Internet of Things
LoRa	Long Range
LPWAN	Low-Power Wide Area Network
LTE	Long Term Evolution
ML	Machine Learning
M2M	Machine to Machine
NB-IoT	Narrowband IoT
NDVI	Normalized Difference Vegetation Index
PM	Particulate Matter
QOS	Quality of Service
RH	Relative Humidity
SSN	Static Sensor Networks
TCP/IP	Transmission Control Protocol/Internet Protocol
VoLTE	Voice over Long Term Evolution
WAN	Wide Area Network
WHO	World Health Organaisation
WiFi	Wireless Fidelity
WSN	Wireless Sensor Network

Chapter 1

Introduction

1.1 Motivation

Outdoor air pollution is a major environmental health problem affecting everyone. Outdoor air pollution in cities and rural areas was estimated to cause 4.2 million premature deaths worldwide per year in 2019; this mortality is due to exposure to air pollutants such as fine particulate matter (PM), CO, NO_2. PM causes cardiovascular and respiratory disease and cancers [1]. The major components of PM are sulfates, nitrates, ammonia, sodium chloride, black carbon, mineral dust [1]. High levels of CO can cause difficulty breathing, tiredness, disorientation, and flu-like symptoms and concern people with some types of heart diseases [2]. Breathing air with a high concentration of NO_2 can irritate airways in the human respiratory system. Exposures over short periods can aggravate respiratory diseases, particularly asthma, leading to respiratory symptoms (such as coughing, wheezing, or difficulty breathing), hospital admissions, and visits to emergency rooms [1].

The lack of spatiotemporal high-resolution exposure data of people living in cities is a major impediment to the extensive epidemiological analysis of the effects of $PM_{2.5}$ [3], CO and NO_2 [4]. Low-cost sensors (LCS), which can be widely deployed at a substantially lower cost than regulatory stations, are a promising compliment for $PM_{2.5}$, CO and NO_2 exposure assessment [5]. When installed in the appropriate locations, LCS improves spatial-temporal understanding of street-level ambient concentrations [6] by measuring air pollutants that vary considerably over short distances due to different emission sources [7]. There are many successful examples of installed fixed LCS networks monitoring air quality and air pollutants [8, 9, 10]. However, fixed LCS networks in large cities still have their limits in terms of geographical coverage, installation permissions, power, and connectivity. An alternative approach is to mount LCS on vehicles that traverse the city, such as on public buses [11], or install it into cars [12] and call these mobile platforms [13]. This term includes mobile sensing in which citizens carry portable handheld IoT devices [14].

If rigorously calibrated, low-cost devices can perform accurately and provide continuous reliable readings of the air quality [15, 16]. However, there is a lack of comprehensive comparative evaluations of the performance of these low-cost sensors in indoor and outdoor mobile environments. This thesis

seeks to address this gap by focusing on the calibration of low-cost sensors using different machine learning (ML) models, not only in controlled indoor environments but also in dynamic outdoor mobile settings. The central objective is to determine whether these sensors exhibit variations in their performance under different conditions and to what extent. Additionally, the thesis aims to identify pollution hot spots within urban areas by deploying the calibrated LCS on a mobile search agent.

1.2 Summary of Contributions

This thesis focuses mainly on the calibration of IoT LCS using ML algorithms. The study presents the importance of calibrating the devices in mobile environment when using them for mobile experiments. The main contributions from this thesis are presented in the chapter 4 are mentioned as follows:

- To calibrate several low-cost air quality sensors, multiple machine learning algorithms were used and their performance was compared to establish the optimal approach for each sensor type. The results revealed that different algorithms performed better for different sensors, emphasizing the necessity of selecting the appropriate algorithm for the individual sensor being used.
- The performance of two low-cost IoT devices is evaluated in mobile and laboratory settings. The need for mobile calibration of IoT devices for mobile air pollution measurement is established.
- A method was proposed for using low-cost IoT devices mounted on a mobile platform to detect $PM_{2.5}$, CO and NO_2 emission spikes in real time. Protocol for using mobile sensing to detect these air pollutant hot spots in urban environments by traversing through the city on different days is presented.

1.3 Structure of Thesis

The rest of this thesis is organized as follows-

- **Chapter 2** provides a quick introduction to IoT and ML, as well as discussions of IoT applications and issues, ML applications, and ML for IoT devices in sensor calibration.
- **Chapter 3** gives an overview of the related work and literature about air pollution monitoring networks and low-cost sensors and calibration of low-cost sensors.
- **Chapter 4** presents calibration results for the two IoT devices. Protocol for using mobile sensing to detect $PM_{2.5}$ CO, and NO_2 pollutant hotspots in urban areas using these calibrated devices deployed on a street car is discussed. The experiments are conducted in Hyderabad, India.
- **Chapter 5** contains the conclusion of this thesis.

Chapter 2

A Brief Overview of IoT and ML

This chapter introduces the IoT and ML. The basic IoT building blocks are discussed, followed by a few IoT-related challenges and a quick rundown of IoT-enabled technologies and applications. ML concepts are introduced, and a few ML applications are discussed, followed by a discussion of ML algorithms for IoT-enabled devices. This chapter is only a brief introduction; interested readers can learn more about IoT and ML from a variety of books and articles, such as [17, 18, 19, 20, 21]

2.1 Internet of Things

IoT is a network of interconnected computing, mechanical, and digital devices that have unique identifiers (UIDs) and can transfer data over a network with little or no human-to-human or human-to-computer interaction. It is the networking of physical objects that contain electronics embedded within their architecture to communicate and sense interactions amongst each other or concerning the external environment [22]. The 'Thing' in IoT can be any device like shown in the Fig.2.1 with the ability to collect and transfer data over a network without manual intervention. The embedded technology in the object helps interact with internal states and the external environment, which helps in decision-making. The concept of IoT is not new, but recent technological advancements, such as miniaturization, low-power sensors, and wireless connectivity, have made it a reality. This has opened up a world of possibilities for businesses and consumers alike, leading to improved efficiency, convenience, and automation of various tasks.

The IoT's ecosystem comprises web-enabled smart devices that use embedded systems, such as sensors, processors, actuators, and communication hardware, to gather, send, and act on the data they gather from their surroundings. By connecting to an IoT gateway, which then sends the data to the cloud for analysis, IoT devices share the sensor data that has been gathered. These devices interact with one another and take action based on the information they exchange. These web-enabled devices' connectivity, networking, and communication protocols are largely determined by the particular IoT applications being used. Fig.2.1 represents the illustration of devices connected showing an IoT network.

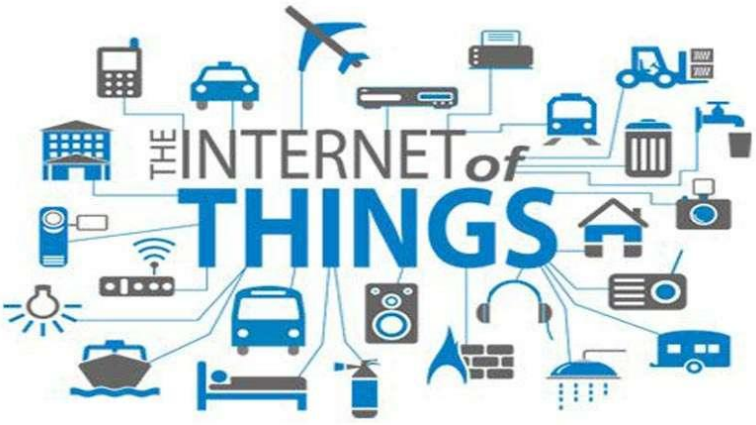


Figure 2.1: Internet of Things [23]

2.1.1 Characteristics

IoT has the potential to revolutionize many aspects of our lives, from improving the efficiency of industrial processes to enhancing the convenience and safety of our homes. Fig.2.2 shows various features of IoT. Here are some key characteristics and benefits of IoT:

- **Connectivity**: IoT relies heavily on connectivity. Devices in the IoT ecosystem are linked to one another and to the internet to enable communication and data exchange. Various technologies such as Wi-Fi, Bluetooth, cellular networks, and satellite communications are used to connect devices. To optimize power consumption and network utilization, IoT devices typically use low-power, low-bandwidth communication protocols. The success of IoT applications is heavily reliant on dependable and secure connectivity. It is critical to ensure proper connectivity for efficient data transmission, real-time monitoring, and analysis of large volumes of data generated by IoT devices. Developing robust and scalable connectivity solutions remains a major challenge as the number of IoT devices grows.

- **Data collection and analysis**: IoT devices generate massive amounts of data, which must be collected and analyzed in order to derive insights and drive intelligent decision-making. IoT data collection can take place at different points in the data lifecycle, such as at the device, gateway, and cloud levels. Machine learning and artificial intelligence are common data analytics techniques used to analyze IoT data and uncover patterns, trends, and anomalies. Insights gained from IoT data analysis can be used to optimize processes, improve efficiencies, and drive innovation in a variety of industries, including healthcare, transportation, and manufacturing

- **Automation and control**: The ability to remotely monitor and control devices via the internet has revolutionized many industries and opened up new avenues for automation. IoT devices can be programmed to respond to specific triggers such as temperature, motion, or light changes and can be used to automate a variety of processes ranging from environmental control to inventory

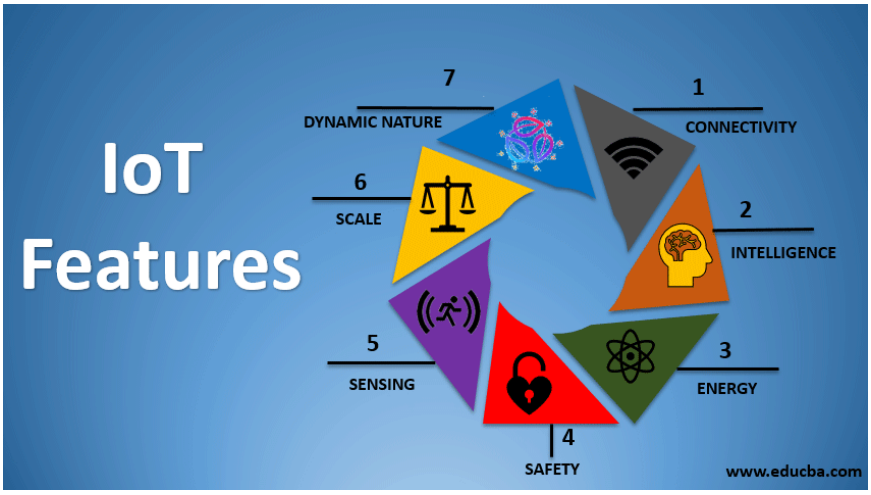


Figure 2.2: Features of IoT [24]

management. By eliminating the need for manual intervention, IoT automation can help improve efficiency, lower costs, and increase safety. Control systems in the Internet of Things can be centralized or decentralized and can be accessed and controlled from any location with an internet connection. Automation and control in IoT are especially useful in industrial settings, where efficient and dependable control systems are critical.

- **Improved safety and security**: IoT has transformed how we interact with our environment, but it has also sparked worries about safety and security. However, technological advancements have led to a substantial enhancement in the security and privacy of IoT devices. Manufacturers are implementing encrypted communication protocols, hardware encryption, and strong authentication to prevent unauthorized access and protect user data. Furthermore, ML algorithms are used to identify and avoid potential threats, whereas software updates can address vulnerabilities that may develop over time. Regulations and standards are also being developed to ensure IoT devices are built with security. All of these advancements ensure that the IoT ecosystem becomes safer and more secure for users, resulting in greater adoption and trust in this technology.

- **Scalability**: The ability to handle a large number of devices and data traffic is critical for ensuring the system's efficiency and reliability. Scalability in the IoT can be achieved through a variety of methods, including cloud computing, edge computing, and distributed computing. Edge computing distributes computing power to devices located closer to the source of data, whereas cloud computing allows for centralized management of devices and data. Furthermore, distributed computing allows multiple devices to collaborate to complete complex tasks, making the system more robust and resilient. With the increasing deployment of IoT devices, scalability is more important than ever, and it is critical to ensure that IoT systems can handle the increasing demand for data processing and storage.

- **Interoperability**: It is a critical feature of IoT, allowing devices and systems from different manufacturers to work together seamlessly. Interoperability in IoT refers to devices' ability to communicate via common communication protocols, data formats, and interfaces. Devices can now exchange data and coordinate actions, allowing for more complex and robust applications. Interoperability is required to develop scalable, flexible, and adaptable IoT systems to changing environments. It also contributes to lowering the cost and complexity of deploying IoT systems by allowing devices and systems to be easily integrated with one another. The development of interoperable standards and protocols is a continuous challenge in the IoT industry, but it is critical for realizing IoT's full potential.

2.1.2 Applications

IoT is a rapidly expanding field with numerous applications in a wide range of industries and domains. Fig. 2.3 shows few of the applications of IoT. The following are some of the most common IoT applications:

- **Smart homes:** IoT devices can be used to automate and control various aspects of a home, including lighting, heating, cooling, security systems, and appliances. [25] examines the use of IoT in smart grid systems and how they work with smart houses. Smart thermostats, such as Nest (now Google Nest) [26], manage heating and cooling based on user preferences and occupancy. Amazon Alexa [27] is speech assistant which operates on a wide range of smart home devices, such as lighting, thermostats, and door locks.

- **Healthcare**: IoT devices can be used to remotely monitor patients, track medication usage, and collect health data to improve treatment outcomes. IoT devices capture and send vital signs and health data to healthcare practitioners, resulting in better patient outcomes and fewer hospital visits. [28] provides an overview of IoT applications in healthcare and analyzes their potential impact on patient care. Smart watches sucg as Apple watch, Fibits [29] track heartrate, sleep pattern, step count calories burned and other health and fitness parameters.

- **Agriculture**: IoT devices can be used to track weather patterns, monitor soil moisture levels, and optimize irrigation and fertilization to increase crop yields. [30] gives an overview of IoT applications in precision agriculture and how they improve agricultural productivity and resource utilization. Microsofts' FarmBeats [31] is a cloud-based platform that helps farmers collect, analyze, and act on data from their farms. It offers farmers a number of tools and features, such as data collection and analysis, as well as actionable insights, resulting in increased crop yields, lower costs, improved efficiency, and increased sustainability.

- **Industrial automation**: IoT devices can be used to track and manage manufacturing, logistics, and supply chain processes. IoT can be used in industrial settings to monitor equipment status, predict repairs, and optimize manufacturing processes. It allows for real-time data analysis to

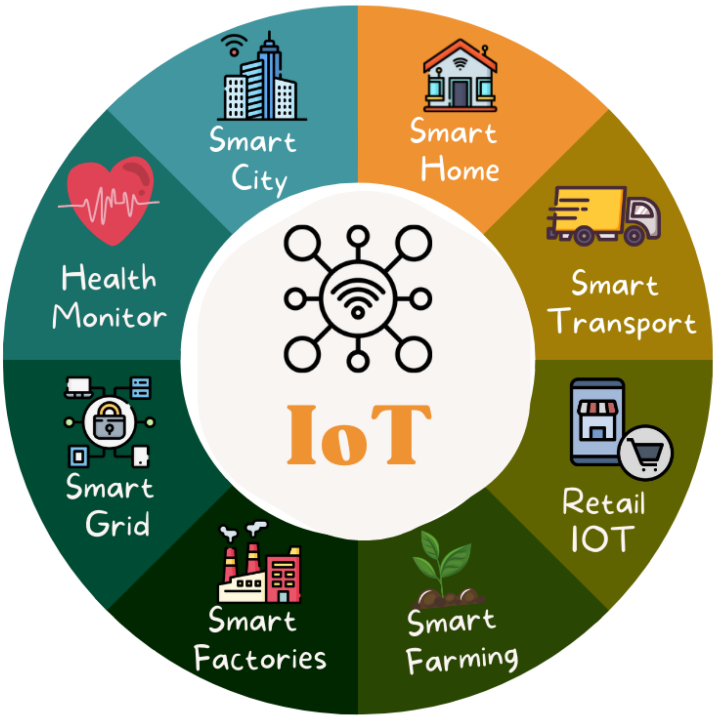


Figure 2.3: Applications of IoT [34]

boost productivity and decrease downtime. Companies like Honeywell, Siemens [32, 33] are offering a variety of IoT solutions for industry, including predictive maintenance, asset tracking, quality control, energy management, logistics and supply chain management, and safety monitoring.

- **Smart cities**: IoT devices can be used to track and control a variety of city systems, including energy use, public transportation, and traffic flow. [35] discusses the utilization of IoT-enabled smart city technologies to keep monitor crowds, enforce social distance, and manage healthcare resources during the COVID-19 epidemic. The Smart City Living Lab [36] at the International Institute of Information Technology Hyderabad (IIITH) is developing and testing smart city technologies. The lab is working on a variety of projects, including air quality monitoring, water quality monitoring, smart energy management, crowd monitoring, [37] shows a novel approach for remote-triggered laboratory experiments using IoT and computer vision. Where several simple lab experiments can be conducted online using IoT. [38] describes a low-cost IoT solution for retrofitting analog water meters with smart capabilities. The solution uses a small, low-power sensor that is attached to the water meter. The sensor collects data on water usage, which is then transmitted to the cloud for analysis. The data can be used to identify leaks, track water usage patterns, and optimize water consumption.The paper demonstrates the feasibility of the IoT solution for retrofitting analog water meters.

- **Environmental monitoring**: To protect natural resources and improve public health, IoT devices can be used to monitor air quality, water quality, and other environmental factors. IoT enables the real-time monitoring of these environmental factors. These sophisticated monitoring systems support effective resource management, early pollution detection, and long-term environmental preservation [30].

2.1.3 Challenges

IoT is a rapidly evolving field with a wide range of applications and use cases, but it also presents a number of challenges that must be addressed in order for its full potential to be realized. Among the key challenges in IoT are:

- **Sensor reliability**: Sensor reliability is a major challenge in IoT implementation. Low-cost, small IoT sensors have limitations that make achieving high data density difficult. Sensor accuracy can be affected by factors such as temperature, humidity, and electromagnetic interference, leading to measurement inaccuracies. Regular calibration of sensors is important to ensure their accuracy and reliability. Sensors can fail due to various reasons, such as wear and tear, environmental factors, or manufacturing defects. Sensor failure can lead to data loss or inaccurate data, which can negatively impact IoT systems.

- **Power consumption**: Power consumption is a significant challenge for many IoT devices because they are battery-powered and have limited computing resources. IoT developers must design energy-efficient devices and systems to extend battery life and reduce environmental impact.

- **Interoperability**: As previously stated, interoperability is critical for developing scalable, flexible, and adaptable IoT systems. Due to the lack of standards and the variety of IoT devices and systems, achieving interoperability can be difficult.

- **Security**: Due to their frequent deployment in unsecured environments and the absence of built-in security features, IoT devices and systems are susceptible to a variety of cyber threats. It is critical to ensure the security of IoT devices and systems in order to protect users and prevent cyber attacks. [39] proposes a number of solutions to mitigate these threats, including end-to-end encryption, protocol and dashboard security, and a deauthentication detector.

- **Data management**: IoT devices generate massive amounts of data, and effectively managing this data is a significant challenge. IoT data is frequently unstructured and originates from a variety of sources, making it difficult to process and analyze.

- **Privacy**: IoT devices and systems frequently collect sensitive data, such as personal health information, and it is critical to ensure the privacy of this data. IoT data collection, storage, and sharing must adhere to privacy regulations and best practices.

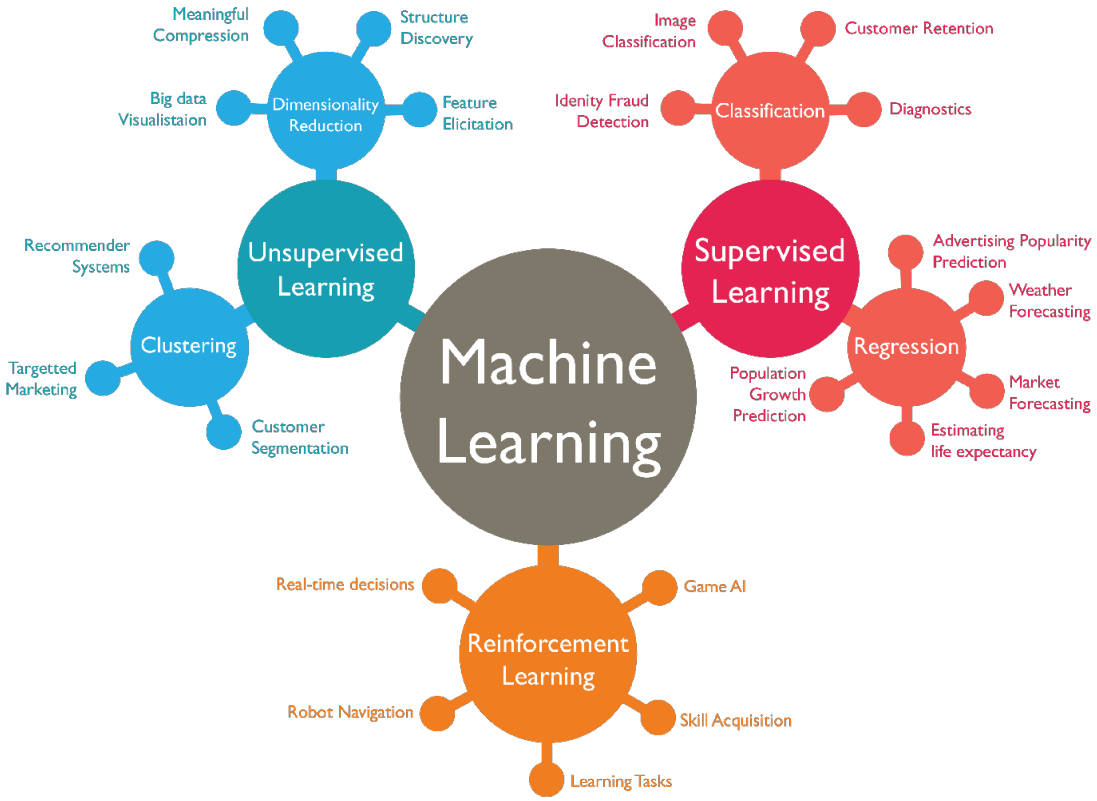


Figure 2.4: Machine Learning [44]

2.2 Machine Learning

ML is a branch of artificial intelligence (AI) that includes the creation of algorithms and statistical models that allow computer systems to learn from data and make predictions or judgments without being explicitly programmed to do so. Because of the increase in data availability and the necessity to automate decision-making processes across a variety of sectors, ML has grown in significance in recent years.

ML is based on the idea that computer systems may learn from data and improve their performance over time by detecting patterns, correlations, and anomalies. The computer system refines its models and predictions based on input from the data in an iterative learning process. This section only provides a brief introduction to ML, more information can be found in [40, 41, 42, 43] .

2.2.1 Categories of ML

The three primary categories of Ml are shown in the Fig. 2.4, and are explained in detail:

- **Supervised Learning**: In supervised learning, each sample in the training data is connected to a goal or output value, and the system is trained using labeled data. The aim of supervised learning

is to learn a mapping function that can forecast an output value given an input variable. The output data in supervised learning is frequently referred to as the target or label, and the input data is frequently referred to as features or predictors [45]. By reducing the discrepancy between the expected and actual results, the algorithm learns to assign labels to the input characteristics. Supervised learning can be further divided into two subcategories:

  - **Regression**: In regression, the objective is to develop a function that can predict the value of the output variable based on the input characteristics. The output value is a continuous variable. Predicting stock prices, property values, or the weather are a few examples of regression problems.
  - **Classification**: Classification aims to train a function that can classify the input features into one of several categories or classes. The output value of classification is a categorical variable. Detecting fraud in credit card transactions, recognizing spam emails, and categorizing photos are a few examples of categorization issues.

  Supervised learning is widely utilized in a variety of applications, including computer vision, natural language processing, and speech recognition. For training, it needs labeled data, which may be time- and money-consuming to acquire. A good model, however, may be used to generate precise predictions on fresh, unobserved data once it has been trained. In this dissertation, various low-cost air quality sensors were calibrated using regression models.

- **Unsupervised Learning**: Clustering, anomaly detection, and dimensionality reduction are the methods employed most frequently in unsupervised learning. Anomaly detection programs find data points that deviate considerably from the rest of the data, whereas clustering algorithms gather together comparable data points based on some similarity measure. The most crucial information is preserved while the number of features in the data is reduced through dimensionality reduction methods. Exploratory data analysis frequently uses unsupervised learning to understand the underlying structure of the data or to preprocess the data before using supervised learning methods. Applications like fraud detection, anomaly detection, and recommendation systems all make use of it.

- **Reinforcement Learning**: Reinforcement learning involves a computer system learning by interacting with its surroundings and getting feedback in the form of rewards or punishments. Learning a policy that maximizes the cumulative reward over time is the aim of reinforcement learning. Several industries, such as robotics, gaming, and autonomous driving, use reinforcement learning. For instance, a reinforcement learning agent in robotics may learn to manipulate things or navigate a space. A reinforcement learning agent can master games like 'chess' or 'go' at a level that rivals a human player's. A reinforcement learning agent can pick up on traffic patterns and collision avoidance while autonomously driving. Environment, agent, and reward function are the three main parts of reinforcement learning algorithms. The states, behaviors, and rewards

are determined by the environment. When the agent engages with its surroundings, it gains the ability to make decisions that will maximize the predicted return. The reward function specifies the agent's objectives, which also offers feedback.

2.2.2 Application of ML

ML has several applications in industries such as healthcare, finance, manufacturing, and marketing, to name a few. These are some examples of ML applications:

- **Image and speech recognition:** In image and speech recognition systems, ML techniques are employed to recognize and categorize objects and speech patterns. Image recognition and computer vision tasks have seen substantial progress thanks to machine learning. Deep Convolutional Neural Networks (CNNs) have excelled at image classification challenges, outperforming conventional techniques. [46] shows the efficacy of deep learning in image recognition, resulting in the widespread use of CNNs in a variety of computer vision applications.

- **Natural language processing:** ML is employed to create systems for understanding, interpreting, and producing human language. Attention mechanisms have improved the performance of language translation, sentiment analysis, and text production tasks in NLP.

- **Fraud detection:** By examining trends and abnormalities in transaction data, ML is used to identify fraudulent actions in the banking, insurance, and e-commerce sectors. [47] addresses the issues of unbalanced data in fraud detection and looked into the use of undersampling to calibrate the probability estimates of fraud detection models, with the goal of boosting the performance of fraud detection systems in real-world financial transactions.

- **Healthcare:** The analysis of medical pictures and data and the development of prediction models for illness diagnosis and therapy all include the application of ML. [48] focuses on the use of machine learning in medical image processing, disease diagnosis, forecasting patient outcomes, and medication discovery. Deep learning algorithms, in particular, have demonstrated promising results in detecting diseases from medical imagery such as X-rays and MRIs. Furthermore, ML models are used to forecast patient readmission rates and risk stratification, allowing healthcare personnel to make better educated decisions and improve patient care.

- **Recommendation systems:** ML algorithms are used to create recommendation systems, which make suggestions to consumers for goods, services, and information based on their preferences and previous actions. [49] introduced collaborative filtering techniques, where matrix factorization was used to model user-item interactions and make personalized recommendations. It has a considerable impact on the development of modern recommendation systems, and collaborative filtering is still a popular strategy in recommendation algorithms.

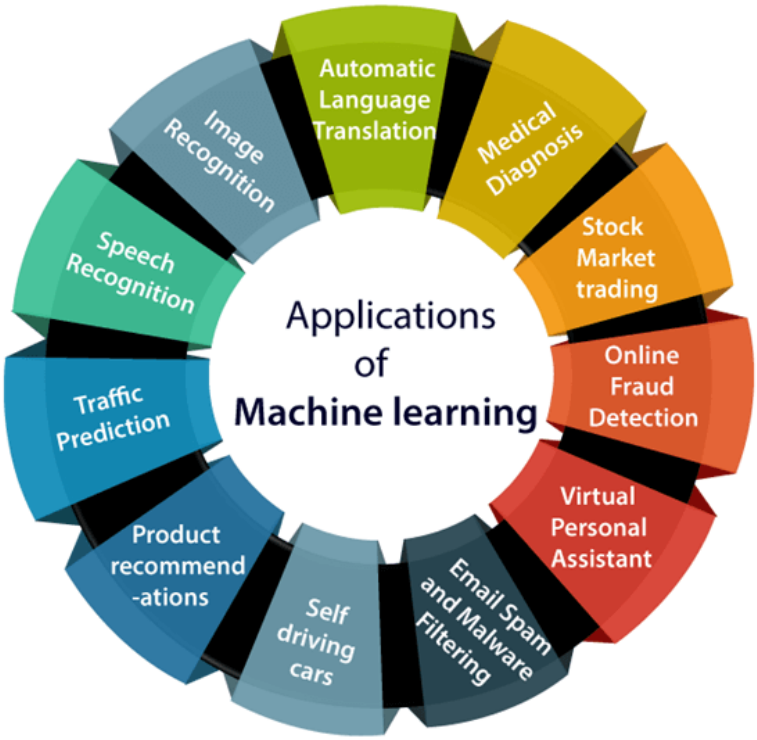


Figure 2.5: Applications of ML [44]

- **Autonomous vehicles:** In self-driving automobiles, ML algorithms are utilized to assess sensor data and make vehicle control choices. [50] presents a deep neural network-based strategy for learning self-driving behaviors from raw sensor inputs from start to finish. This trailblazing effort has set the road for advances in autonomous vehicle technology, bringing us closer to the realization of safe and dependable self-driving automobiles in the future.

The discipline of ML is expanding quickly and has the potential to completely change how data is processed and analyzed. ML is anticipated to continue playing a significant role in many sectors due to the expansion of data availability and the creation of new algorithms and models.

Chapter 3

Overview of IoT Based Air Quality Monitoring Networks and Low-Cost Sensors

3.1 Introduction

This chapter briefly outlines the reason for working on dense air pollution monitoring. A complete literature study of previous approaches and traditional monitoring sensor networks, low-cost sensors (LCS) for air pollution monitoring, and a thorough survey of various current IoT air pollution monitoring networks worldwide are briefly covered

3.2 Air Pollution Monitoring Network

3.2.1 Stationary Network

The atmospheric concentration of $PM_{2.5}$, CO, and NO_2 measured by fixed monitoring stations equipped with certified reference instruments are often sparsely deployed throughout the city due to their high cost (INR 1.5-2.5 crores) [51]. Governmental agencies often deploy and maintain these networks to keep the public informed about the AQI. This sparse deployment leads to low spatial resolution and misrepresentation of street-level concentrations [6] because air pollutants vary considerably over short distances due to different emission sources [7]. [52] study looks on the impact of monitoring station placement on urban air quality assessment and underlines the importance of optimal placement in recording fluctuations in pollution levels and directing effective pollution control strategies. It is commonly suggested that large-scale deployment of LCS with a high spatial-temporal resolution with real-time access to pollution data helps to resolve this issue [5]. Accordingly, there are many examples of dense deployment of fixed low-cost devices in cities, such as Hyderabad, India [8], Cambridge, UK [9], and California, USA [10]. [53] explores the use of low-cost IoT sensors to improve the spatial and temporal understanding of particulate matter (PM) pollution; authors deployed a network of nine low-cost IoT sensors in a small educational campus in Hyderabad, India. In [8], 49 IoT devices were densely deployed in an area of

$4\,\mathrm{km}^2$ in Gachibowli, Hyderabad, India. The data from the devices can be viewed on a webpage and an android app as the extension of the previous work in [53]. In [54], 33 measuring units located around London. A few more nodes put by the local authorities offer additional data to the network. This information is freely available to the public, and anybody may use the website to monitor the air quality in real time. [55] project attempts to create a wireless sensor network with the coverage of a Cambridge. A total of 100 Linux-based PCs are installed in diverse locations such as streetlamp posts and poles. The nodes are outfitted with radios that function as a mesh network. Data is continually posted to servers and made available to the public via a web app.

3.2.2 Mobile Network

Less common is the application of mobile low-cost sensor monitoring to evaluate the spatial heterogeneity of air pollutants and avoid the costs of power and connectivity. In mobile low-cost sensor air quality monitoring, IoT equipment is mounted on a mobile platform or handheld. Citizens attaching the lightweight, portable handheld device on their backpacks collecting their exposure data in Helsinki [14]. GasMobile [56] is a gadget created by researchers that can detect outside air pollution and connect directly to a smartphone through a USB port.

Google Street View automobiles, equipped with high-frequency lab-graded air pollutant monitoring devices and tested in different cities [7]. Sensors are installed in the Palermo bus fleet in Italy [11]. Devices fixed into cars providing long-term mobile datasets of air contaminants in Ontario, Canada [12]. [57] describes a mobile air quality monitoring system that uses a Raspberry Pi as the main controller. The system provides real-time air quality data for a specific location, which can be used by individuals to assess their exposure to air pollution. [58] proposes a mobile sensing platform for smart city services. The platform, called City Scanner, is a modular system that can be used to collect data on a variety of city features, such as air quality, traffic, and noise pollution. [59] uses a fleet of vehicles equipped with sensors to measure air quality parameters. The vehicles are driven along predetermined routes in 3 cities of California, and the data collected by the sensors is used to create maps of air pollution levels. [60] presents a methodology for mapping spatial variation of air pollution levels in the city of Antwerp, Belgium. [61] conducted a mobile air quality monitoring study in Sydney, Australia along a busy roadside location in the suburb of Randwick. [62] proposes approach for estimating the AQI using image processing and learning methods. However, more research is needed to validate the proposed approach on larger datasets and in different environments.

3.3 Low-Cost Sensors

3.3.1 Particulate Matter

Tiny particles in the air are referred to as particulate matter (PM). The elements that make up these particles can include a wide range of things, including dust, dirt, soot, smoke, and liquid droplets. Ac-

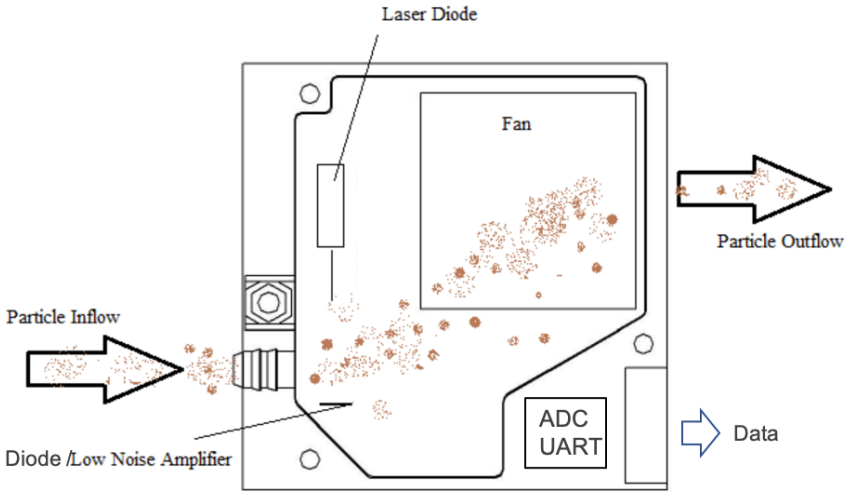


Figure 3.1: Optical scattering principle in PM sensors [63]

cording to its size, PM is categorized, with $PM_{2.5}$ and PM_{10} being the most often measured categories. Particles with a diameter of 2.5 micrometers or less are referred to as $PM_{2.5}$, while those with a diameter of 10 micrometers or less are referred to as PM_{10}. PM can be hazardous to human health, especially when inhaled. Long-term exposure to high amounts of PM has been associated with an increased risk of heart disease, stroke, and lung cancer. It can exacerbate respiratory and cardiovascular diseases. Moreover, PM can have an adverse impact on the environment, harming crops and ecosystems, impairing visibility, and altering the planet's temperature.

There are different types of low-cost PM sensors available in the market. These PM sensors use various working principles to estimate the concentration of airborne particles. The LCS used in this project use optical scattering method. These sensors transmit light into the air using a light source, similar to an LED. A photodetector detects the intensity of the dispersed light after it is scattered by airborne particles. Estimating PM levels is possible because to the relationship between the intensity of scattered light and particle concentration in the air. The working principle is demonstrated in the Fig.3.1 and the sensor details are explained below:

- **Prana Air:** PM, dust particles are measured using an optically built industrial-grade, digital laser sensor. It has a laser and a photoelectric receiving module. It operates on the 90° light scattering principle. Light that strikes the mirror's aperture at 90° is reflected towards the sensor. For as long as the light is reflected, the photodiode registers a pulse. The electrical signal thus received is converted into the concentration of PM by specific algorithms [64].

- **SPS30:** The Sensirion SPS30 operates on the laser scattering principle. A fan creates a controlled airflow inside the sensor. Environmental PM is transferred by the airflow inside the sensor from input to output. Light scattering occurs when particles in the airstream pass across a focused laser beam in line with the photodiode, Sensirion's unique algorithms, which run on the SPS30 internal

microcontroller, detect the scattered light and convert it to a mass/number concentration output [65].

3.3.2 Gas Sensors

Gas sensors detect and quantify the concentration of various gases in the air or in a specific environment. Gas sensors are classified into several categories, each of which is designed to detect different gases using distinct detecting technologies. This project studies about the CO and NO_2 gas sensors.

- **Carbon monoxide:** The incomplete combustion of fossil fuels, including coal, natural gas, propane, and gasoline, results in the production of carbon monoxide (CO), an odorless, tasteless, and colorless gas. Additionally, wood-burning stoves and cigarette smoke also contribute to its production. Most outdoor CO emissions to ambient air come from vehicle exhaust [66]. Exposure to CO can cause difficulty in breathing, tiredness, disorientation, and other flu-like symptoms. CO exposure at extremely high concentrations can be fatal [1]. Headache, dizziness, weakness, nausea, vomiting, disorientation, and loss of consciousness are all signs of carbon monoxide overdose. Long-term health issues like heart disease, neurological damage, and cognitive impairment can also result from prolonged exposure to low levels of carbon monoxide.

- **Nitrogen dioxide:** Nitrogen dioxide (NO_2) is a dangerous air pollutant that is released from sources like industrial activities, power plants, and vehicle exhaust. Exposure to excessive concentrations of NO_2 can have a range of harmful health impacts on people and animals. It is a reddish-brown gas with a strong stench. NO_2 is an indicator for calculating and evaluating air pollution from motor vehicle sources [67]. NO_2 is linked to respiratory disorders, notably asthma, since it irritates the lungs and exacerbates respiratory ailments [1]. Moreover, NO_2 may combine with other airborne molecules to generate hazardous PM, which can worsen respiratory conditions and increase the risk of cardiovascular disease.

The gas sensors used and their working principles are explained in detail below:

- **MICS-4514:** The MiCS-4514 is a compact metal oxide semiconductor (MOS) sensor that contains two totally independent sensing elements in a single package [68]. It is a robust micro electro mechanical sensors that can measure CO and NO_2. The silicon gas sensor construction is made up of a precisely micro machined diaphragm with an embedded heating resistor and a sensing layer on top. Two sensor chips with independent heaters and sensitive layers are included in the MiCS-4514. The first sensor chip detects oxidising gases (OX), whereas the second detects reducing gases (RED). The MICS-4514 measurement circuit is show in the Fig. 3.2. On each sensor, constant power is the preferred mode of operation. The RED sensor has a nominal power of PH = 76 mW, whereas the OX sensor has a nominal power of PH = 43 mW. The resultant sensor layer temperatures are roughly 340 °C and 220 °C in air at approximately 20 °C The pollutant gases are detected by measuring the sensing resistance of both sensors. In the presence

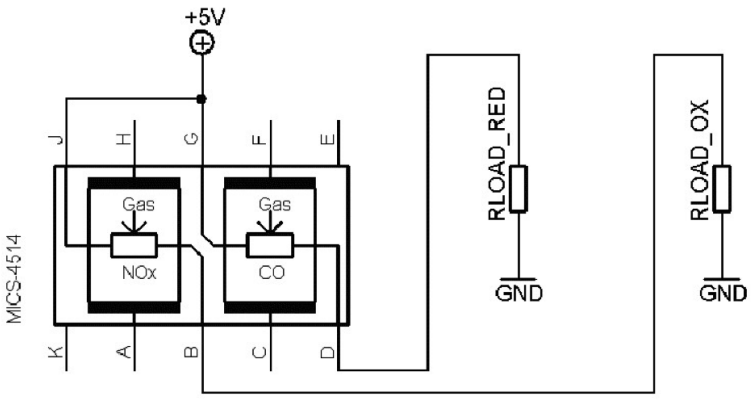


Figure 3.2: MICS-4514 circuit [68]

of CO and hydrocarbons, RED sensor resistance reduces. In the presence of NO_2, the resistance of the OX sensor rises.

- **SPEC CO & NO_2:** SPEC sensors are amperometric gas sensors, which are electrochemical sensors that generate a current proportional to the volumetric fraction of the gas. In the Fig. 3.3 the two electrodes are shown in contact with a liquid electrolyte in a conventional electrochemical sensor [69]. The gas is measured at the working electrode, which is typically a catalytic metal chosen to optimize the target gas's reaction. The measured gas enters the capillary diffusion barrier and reacts with the electrode. The electrons produced by the electrochemical reaction flow to or from the working electrode via an external circuit, depending on the amount of gas reacting. The sensor's output signal is the working electrode current. The counter electrode serves only as the second half-cell, allowing electrons to enter and exit the electrolyte in equal numbers and in the opposite direction as those involved in the working electrode reaction. The reference electrode, which creates a steady electrochemical potential in the electrolyte that is often shielded from exposure to the sample gas, enhances the stability, signal-to-noise ratio, and response time of the two-electrode design.

3.4 Calibration of LCS

Calibration is the process of adjusting sensor values to match a known reference, often a laboratory-grade device. If the sensors are not calibrated, they may generate erroneous or inconsistent results, leading to false conclusions regarding air quality. Inaccurate or inconsistent readings can produce data that is ineffective for decision-making or policy creation. Calibrated sensors give reliable and accurate data, allowing you to discover patterns, set baselines, and analyze progress over time. Regular calibration is crucial for detecting and fixing any alterations in sensor performance over time as well as guaranteeing the accuracy and dependability of the data that these sensors collect.

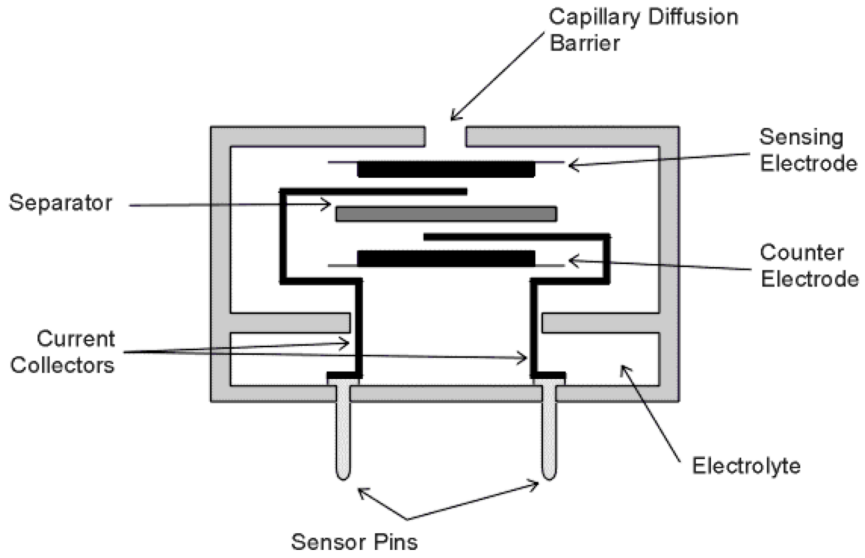


Figure 3.3: SPEC two electrode electrochemical gas sensor [69]

Span calibration is usually performed at the factory immediately after the manufacture of the gas sensor[70]. In span calibration, the LCS are installed in the gas chamber during the calibration, and the known concentration of the gas is circulated through the gas chamber. The LCS' response to changes in gas concentration is monitored, and the LCS are calibrated [71]. In reference-based calibration, the LCS are co-located with the reference grade instrument for an extended length of time and calibrated using ML models. Span calibration ensures that the actual gas measurement is accurate, usually performed right after a gas sensor is manufactured at the factory. When the low-cost sensor is used for mobile measurements, it is necessary to calibrate them in the mobile environment. A portable reference sensor can be used to calibrate the LCS by reference-based calibration method.

This study mainly focuses on the mobile calibration of low-cost gas sensor devices for non-static air quality measurements. The performance of the two low-cost devices is examined by comparing their data with the reference device data in both lab (as shown in Fig.3.4) and outdoor settings. The devices are then calibrated using different ML algorithms, and the effectiveness of these algorithms is also examined in both lab and outdoor settings.

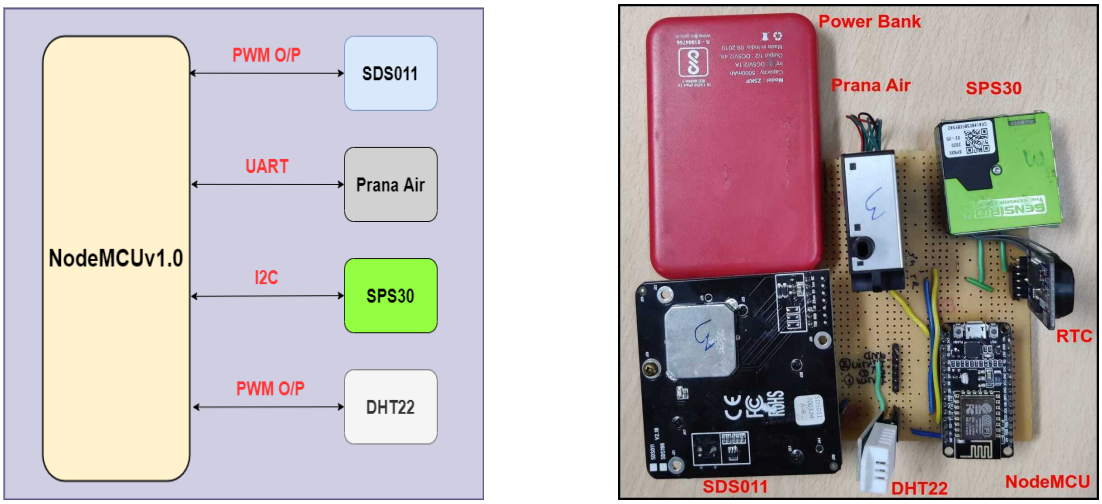
The general trend today is to calibrate low-cost air quality sensors against reference data sets using ML algorithms. For instance, simple linear regression (SLR) and multi-variant linear regression (MLR) were used to calibrate of LCS in USA [72]. SLR, MLR, and artificial neural networks (ANN) were applied to calibrate the sensors measuring O_3, NO, NO_2, CO, and CO_2 in Italy [73]. Similarly, low-cost gas sensors were co-located against reference sensor and calibrated by multiple calibration models like SLR, MLR, random forest regression (RFR), long short-term memory (LSTM), and generalized additive models (GAM) in Beijing [74], and Sheffied [75]. LCS are often moved after calibration. In Pittsburgh, USA 70 sensors were first deployed for a month on the CMU campus with a reference sensor for calibration and then deployed in Pittsburgh city [76]. In China, a 4-stage calibration model

Figure 3.4: Reference based calibration in a gas chamber

was performed, and in the last stage, sensors were calibrated by a mobile reference device using a simple linear model [71].

In the paper [77] calibration of three low-cost sensors SDS011, Prana Air, and SPS30 namely were performed and their performances were compared in both indoor and outdoor environments. Three identical test nodes were created for the experiments. Fig. 3.5 shows the schematic view and actual view, respectively, for each such node, which consists of one unit of SDS011, Prana Air, and SPS30 each. ESP8266 based Wi-Fi enabled NodeMCU v1.0 microcontroller module was used to interface these sensors. Samples were collected at 2 sec intervals, and all data was pushed to Thingspeak, an MQTT-based IoT platform. The performance evaluation of these sensors is carried out in terms of coefficient of determination (R^2), coefficient of variation (C_v), and root mean square error (RMSE). The sensors were calibrated using linear regression. Once calibrated it was observed that the sensors performance is improved in both indoor and outdoor experiments. In both indoor and outdoor experiments the low-cost devices were co-located with the reference device Aeroqual. It should be noted that in the outdoor experiments the the data was collected at few fixed locations. A similar experiment was conducted in [78]; different low-cost CO_2 sensors were compared for indoor air quality monitoring.

However, there are problems with the current approaches to calibrating LCS. Most of the calibration is conducted in a laboratory or co-located against an outdoor stationary reference sensor, or station. In controlled laboratory settings, it is difficult to effectively transform the raw sensor responses into concentration estimates using SLR or other ML models [79]. In sparsely located stationary reference



(a) Block architecture

(b) Actual hardware

Figure 3.5: Hardware setup

stations, there is a limitation that the co-located LCS will never get exposed to all ranges of gases present at the city street level. To fill in this gap, this research suggests that LCS should be calibrated using ML techniques in their deployed environment, and for mobile low-cost sensing, this means the reference device should also be mobile. No such studies have been done on low-cost sensor mobile calibration, with the exception of applying a linear model for calibration. The alternative to this research approach is to gather block-by-block pollution concentrations using Google Street View automobiles equipped with high-frequency lab-graded air pollutant monitoring devices. However, the cost and required logistics for Google cars are not appropriate for many cities. In this research, we mount low-cost air quality sensors and lab-grade reference devices on the local street cars and calibrate them using different ML models in laboratory and block-by-block spatial scales.

Chapter 4

Low-cost $PM_{2.5}$, CO and NO_2 Sensor Evaluation and Calibration for Mobile Platform

This chapter provides the motivation for ML-based calibration of the low-cost air quality sensor for mobile measurements. monitoring air pollution and using IoT as an enabler for it, followed by global initiatives around the world for tackling air pollution, conventional monitoring sensor networks, low-cost sensors for air pollution monitoring, and a thorough survey of various existing IoT air pollution monitoring networks around the world.

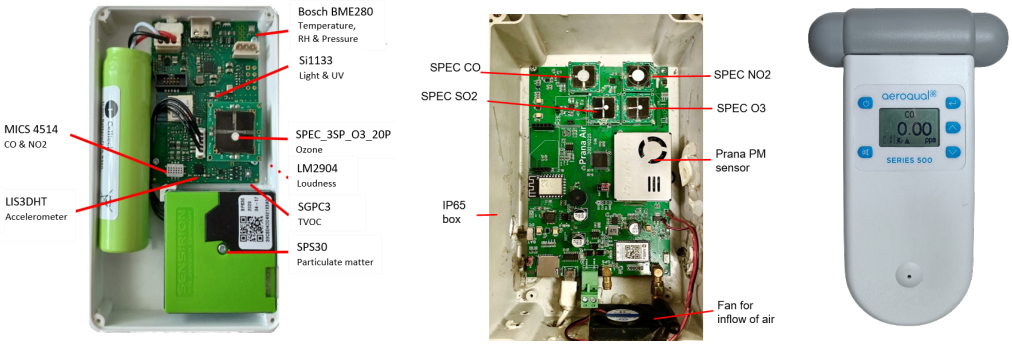
4.1 Introduction

Our contribution to the field of sensor studies are:

- Comparison of the performance of two types of portable, light-weight low-cost sensors to measure short-term fluctuations of $PM_{2.5}$, CO, and NO_2 atmospheric concentration against a reference device in the laboratory (in a chamber) and on a mobile platform (sampling outdoor air quality placed on the roof of a street car).
- Step-by-step evaluation of different ML algorithms to reduce the mean average error between the portable low-sensor types against a portable reference device for both laboratory and mobile platform settings.
- Applying the ML data calibration to block-by-block street car air quality measurements without the need for the mobile reference device and identifying the spatial concentration of $PM_{2.5}$, CO, and NO_2 during Diwali festival week in Hyderabad, India

4.2 Hardware Specifications

The performances of two low-cost types, MegaSense One (Fig. 4.1(a)) and Prana Air (Fig. 4.1(b)), were compared. MegaSense One was developed by the Department of Computer Science and Atmo-



(a) MegaSense sensor type. (b) Prana Air sensor type. (c) Aeroqual reference.

Figure 4.1: Sensors types and components used for the experiments.

spheric Sciences at the University of Helsinki, Finland. Aeroqual series-500 (Fig .4.1(c)) was used as the reference instrument to evaluate the accuracy of low-cost sensors. The details of all the components mentioned above, including the reference instrument, are provided in the following subsections.

4.2.1 MegaSense One

MegaSense One (Fig.4.1(a)) is a portable sensing platform based on a BMD-340 system module and mobile phone app called MegaSense. The platform connects to COTS Android smartphones over Bluetooth Low Energy (BLE), and the smartphone uploads the readings to the MegaSense cloud. The platform consists of sensing components purchased from different manufacturers (Table 4.1), including Bosch BME-280 (temperature, humidity and pressure), Sensirion SPS30 (PM1, 2.5, 4, 10), MICS 4514 (NO_2 and CO), SPEC Sensors LLC 110-406 (O_3). Other accessory components are Sensirion SGPC3 (TVOC), Silabs SI1133 (UV Index and ambient light), STMicroelectronics LIS3DH (acceleration and orientation), and Texas Instruments LM2904 (loudness). The component parameters are measured every 30 seconds, and transmitted in a 350 bytes data packet. The platform is powered with a 3500 mAh battery and enclosed in a 3D-printed case made of ESD-PETG filament. The sensor form dimensions are width 70 mm, depth 25 mm, height 125 mm, and weight 155 grams. The front is protected by an aluminum mesh. The general battery life before recharging via micro USB interface is 26 hours. Indicator LEDs are used for communication and charging.

4.2.1.1 Prana Air Device

Prana (Fig. 4.1(b)) is a portable device that measures the concentration of PM_{10} and $PM_{2.5}$, SO_2, NO_2, O_3, H_2S and CO. Table 4.1 shows the specifications of the sensing components in the device. The following sensing components are used to measure the pollutants - Prana air PM sensor, dedicated SPEC sensors for each gas SO_2, NO_2, O_3, H_2S and CO. VOCs are measured using MICS sensor, and

Table 4.1: Specifications of sensors.

Sensor Type	Components	Parameter	Range
MegaSense	BME 280	Temperature	-40°to 85°C
		Humidity	0 to 100%
		Air Pressure	300–1100 hPa
	SPS30	$PM_{2.5}$	0-999 ppm
	MiCS4514	CO	0-1000ppm
		NO_2	0.05-10 ppm
Prana Air	SHT30	Temperature	0 °to 65 °C
		Humidity	10 to 90%
	Prana	$PM_{2.5}$	0-999 ppm
	SPEC	CO	0-20 ppm
		NO_2	0-20 ppm
Aeroqual		$PM_{2.5}$	0-999 ppm
		CO	0-25 ppm
		NO_2	0-1 ppm

SHT22 is used for measuring temperature and humidity. It is mentioned in the specification sheet that the data is calibrated and can be treated as the actual concentration value. The data is shown on a digital screen that comes with the device, and there is a service that provides access to analytic results. The real-time data is given in less than 30 seconds with accuracy and precision and has Wi-Fi, GPRS, and RS-485 types of connectivity. The data is easily accessible from the AQI website and mobile app.

4.2.2 Reference Instrument

Aeroqual series-500 (Fig. 4.1(c)) is a portable pollution monitoring instrument that logs data in the CSV format at a minimum of 1 min intervals. It comes with several sensor heads, and each head measures a specific pollutant. Only one head can be used at a time. For this experiment, two Aeroqual monitors were used to measure the concentration of CO and NO_2. The reference instrument has the ability to log data over different averaging intervals ranging from 1 min up to 1 hour. The data was downloaded via a USB interface. The specifications of the sensor heads available with Aeroqual have

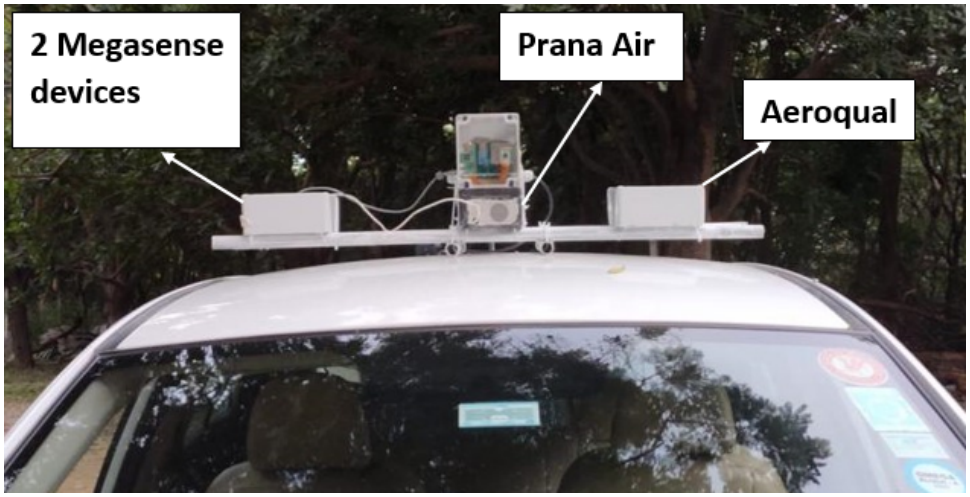


Figure 4.2: Low-cost sensor types and reference instrument on a mobile platform on a street car for data collection.

listed in Table 4.1. Additionally, the official manual [80] reports institutions like NASA, Samsung, Tesla, and others as clients.

4.2.3 Experimental Setup and Measurements

The low-cost sensors were used in three experiments. The first two experiments concerned laboratory and mobile calibrations of sensor measurements against the reference device measurements. The third experiment evaluated CO and NO_2 measurements during the festival of Diwali using the lab and mobile-based calibration models.

The following subsections describe laboratory and mobile experimental setups in detail.

4.2.3.1 Laboratory Experiment

During the $PM_{2.5}$, CO and NO_2 laboratory experiments, both low-cost sensor types and the reference device were placed in a controlled environment. The low-cost sensors were co-located with the reference device in a chamber. Incense sticks were used as a pollution source. Burning an incense stick (the composition of incense stick is not taken into consideration) released a number of pollutants [81]. The incense stick was lit until the reference device started showing a value corresponding to the maximum rating of gas sensors. After that, the smoke was allowed to leak out of the chamber slowly for a few hours. The readings were sampled every 1 sec and then time-averaged to 1 min intervals because the reference instrument did not log at a rate lower than it. The experiments were conducted for approximately 2 and a half days in April (2022), and around 3000 data points were logged. Upon analysis, it was observed that the incense produced large amounts of $PM_{2.5}$ and CO, but showed little or no effect on NO_2 levels.

4.2.3.2 Mobile Platform Experiment

During the $PM_{2.5}$, CO and NO_2 mobile experiments, the low-cost sensors and the reference instrument were mounted on top of a street car, as shown in Fig. 4.2. The car was driven to the different parts of the city for one week each in Nov & Dec. 2021 and Jan. 2022. The data collection path incorporated a variety of settings, including urban, semi-urban, industrial, construction sites, villages, marketplaces, high-traffic areas, and highways. The goal of the data campaign was to evaluate the capabilities of low-cost sensors to detect changes in $PM_{2.5}$, CO and NO_2 levels when subjected to dynamic conditions. The sampling and averaging intervals were the same as the indoor experiment. The wind speed and speed of the vehicle were not taken into consideration. The average speed of the car was around 40 Kmph. The data was collected for 3-4 hours every day during this campaign. Around 4500 data points were collected and used for training the model for mobile calibration of devices. Every day, almost 100 Km were traversed. Around 6800 data points were collected from each device, out of which around 4500 points (66.67%) were used for training the model for mobile calibration of devices.

4.3 Data Processing Methods and ML Algorithms

4.3.1 Data Cleaning, Pre-processing and Data Matrix Definitions

To convert raw data collected from the low-cost sensor types into a usable dataset to following tasks were done:

- The low-cost sensor devices did not send the data at the same time and there were few seconds of deviations within each. We applied data averaging to look past the random changes and fluctuations and to see the major trend of the dataset. The data points were averaged into a single timestamp within every minute time frame giving a dataset with 1-minute sampled data points to correspond with the target variable data from the reference instrument aeroqual.
- The outliers were removed using the standard deviation method before applying ML algorithms.
- 10-fold cross-validation is applied in both laboratory and mobile calibration processes.
- For the following sections, the dataset considered is expressed by the matrix $\mathbf{X}$ of dimensions $M \times N$ where N represents the number of data points obtained for that particular device, and the combination of input features and M represents the number of input features considered and $\mathbf{y}$ is the aeroqual data with the dimension $1 \times N$ which acts as the target labels.

4.3.2 Machine Learning Algorithms

Supervised learning [82] algorithms were applied to calibrate the low-cost sensors' data with respect to the data from a more accurate reference sensor. The data from low-cost gas sensors and other parameters such as temperature, RH, pressure, etc., were used as input variables, and the data from the

more accurate reference sensor were considered the target variable. Regression algorithms were used to estimate the relationship between the input and target variables and reduce the error between estimated values and the target variable. For the following sections, the dataset considered is expressed by the matrix $\mathbf{X}$ of dimensions $M \times N$ where N represents the number of data points obtained for that particular device, and the combination of input features and M represents the number of input features considered and $\mathbf{y}$ is the aeroqual data with the dimension $1 \times N$ which acts as the target variable. The algorithms considered in our experiments are explained in the following subsections:

4.3.2.1 Linear Regression

Linear regression (LR) is one of the easiest regression algorithms with a simple representation. The representational model is a simple linear equation combining input variables to give an output solution as an estimate for the target variable. Specifically, the estimates of the target variable are calculated as a linear combination of the input variables. Depending on the number of input variables considered, LR was classified as Simple (SLR) or Multiple (MLR). For SLR, the input matrix $\mathbf{X}$ was the size $1 \times N$ for each device, as only a single feature is used, which is the pre-calibrated pollutant itself. For example, while calibrating CO values, only the CO data from the low-cost devices were taken as the feature.

For MLR, the input matrix $\mathbf{X}$ with the size $N \times M$ for each device and considering the M set of input features. While using MLR, different combinations of features such as temperature, RH, and pressure were considered. The M value was varied according to the number of features. For example, while calibrating CO, when temperature, and RH, were also taken into account along with the CO data from the device, the value of M is 3.

4.3.2.2 Polynomial Regression

Polynomial regression is an extension of LR models. In simple terms, due to the non-linear relationship between the input and the target variables, polynomial terms were added to linear regression to convert it into polynomial regression. In this case, a 3rd degree polynomial model was trained and tested using the performance metrics. Similar to the case of linear regression algorithms, polynomial regression algorithms can also be classified into simple polynomial regression and multivariate polynomial regression. For simple polynomial regression, the input matrix $\mathbf{X}$ was the size $1 \times N$ for each device, as only a single feature is used i.e., only the raw value of pollutant measured by the low-cost device. For multivariate polynomial regression, the input matrix $\mathbf{X}$ was the size $N \times M$ for each device and considering M set of input features. Different combinations of features were taken into consideration in this case. The M value is varied with the number of features taken into account while running the model.

4.3.2.3 Support Vector Regression

Support vector regression (SVR) is a supervised learning algorithm that is used to predict discrete values. Support vector machines are well known for solving classification problems. However, the use of SVMs in regression is not as well considered. These types of models are known as support vector regression models. SVR uses the same principle as the SVMs. In SVR, the best fit was the hyperplane that had the maximum number of points. The hyperplane is a separating boundary between two data classes in Support Vector Machine (SVM). In SVR, we used this approach to predict the continuous output. We applied the radial basis function (RBF) as the kernel parameter. The SVR was performed considering different combinations of features, such as temperature, RH, and pressure along with the pollutant $PM_{2.5}$ or CO or NO_2. Unlike other Regression models that try to minimize the error between the real and predicted value, the SVR tries to fit the best line within a threshold value. The threshold value is the distance between the hyperplane and boundary lines. These are the two lines that are drawn around the hyperplane at a distance of ϵ.

4.3.2.4 Decision Trees and Random Forest Regression

Decision trees (DT) are models that use a set of binary rules to calculate a target value. Each individual tree has branches, nodes, and leaves. A decision tree arrives at an estimate by trying to ask a series of questions to the input data until the model is confident enough to make a single prediction determined by the model itself during training. Decision tree regression normally uses mean squared error (MSE) to decide to split a node into two or more sub-nodes. Suppose we are doing a binary tree the algorithm first will pick a value and split the data into two subsets. For each subset, it will calculate the MSE separately between the predictions(average values in both subsets) obtained and the target value. The tree chooses the value with results in the smallest MSE value. For predictions, given a data point, it is run through the entire tree up until it reaches a leaf node. The final prediction is the average of the value of the dependent variable in that leaf node. Decision tree regressors are prone to the problem of overfitting. The decision tree can recursively split the data set into a large number of subsets to the point where a set contains only one row or record to reduce the MSE. Even though this might reduce the MSE to zero, this is obviously not a good thing as it overfits the training dataset. Random forest regression (RFR) was the better choice in this case. The DT and RFR were conducted while taking into account various combinations of features such as temperature, RH, pressure, and the pollutant $PM_{2.5}$ or CO or NO_2.

The bootstrapping random forest algorithm combines ensemble learning methods with the decision tree framework to create multiple randomly drawn decision trees from the data, averaging the results to output a new result that often leads to strong predictions. Ensemble learning is the process of using multiple models, trained over the same data, averaging the results of each model, and ultimately finding a more powerful predictive result. Bootstrapping is the process of randomly sampling subsets of a dataset over a given number of iterations and a given number of variables.

4.3.3 Performance Metrics

The comparison criteria considered for analyzing the algorithms' performance for calibration are R^2, MAE, MSE, and RMSE. The performance parameters are explained below:

- **R^2**, also known as the Determination Coefficient, measures how much prediction error is eliminated. It is a statistical measure representing the proportion of the variance for a dependent variable explained by an independent variable or variables during a regression model.

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}, \tag{4.1}$$

where SS_{res} is the sum of squares of residuals and SS_{tot} is total sum of squares, they are given by

$$SS_{\text{tot}} = \sum_{i=1}^{n} (y_{\text{i}} - \bar{y})^2, \tag{4.2}$$

where *n* is the number of samples $\bar{y}$ is the mean of the target data and

$$SS_{\text{res}} = \sum_{i=1}^{n} (y_{\text{i}} - \hat{y}_{\text{i}})^2, \tag{4.3}$$

where $\hat{y}_{\text{i}}$ is predicted value.

- **Root Mean Square (RMSE)**, one of the most commonly used errors metric for evaluating the performance of regression models, gives how much the predicted results differ from the actual value. The equation for RMSE is given by

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_{\text{i}} - \hat{y}_{\text{i}})^2}. \tag{4.4}$$

- **Mean Absolute Error (MAE)** is a statistical measure that assesses the average magnitude of errors in a group of predicted values without taking into account their direction.

  and is given as

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_{\text{i}} - \hat{y}_{\text{i}}|. \tag{4.5}$$

4.4 Results

The results are presented in three steps. The first step analyzes the $PM_{2.5}$, CO, and NO_2 raw data collected from the three low-cost sensors in the laboratory and mobile experiments to assess the need for local calibration. The second step presents lab and mobile-based calibration results using different ML models discussed in Section 4.3. Finally, the mobile data collected is analyzed for the device which performs the best in mobile calibration. For analyzing the results, well-known performance parameters are considered: R^2, root mean square error (RMSE), and mean absolute error (MAE).

Table 4.2: Comparison of performance for raw observations.

	$PM_{2.5}$		CO		NO_2	
Sensor Type	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
	Laboratory experiment					
MegaSense-1	10.18	16.09	277.52	285.05	1169.98	1182.63
Prana	39.52	44.94	0.90	1.20	43.28	48.26
	Mobile experiment					
MegaSense-1	32.54	38.90	291.75	298.97	967.82	1150.77
Prana	15.88	19.48	3.63	5.04	152.31	168.74

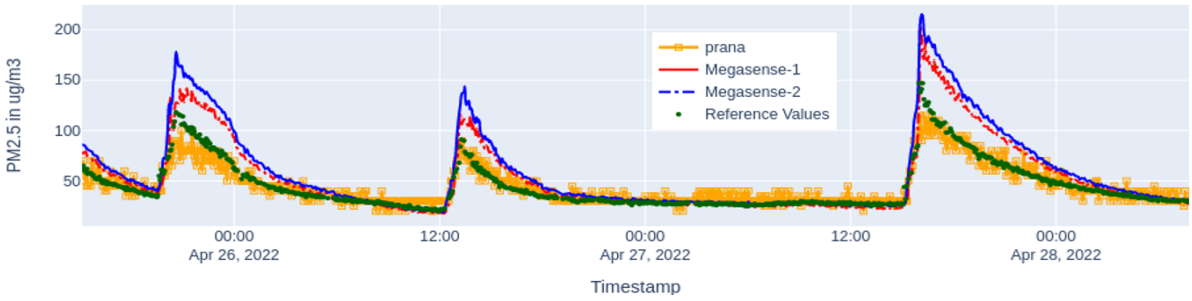
4.4.1 Raw Data Analysis

The low-cost sensor measurements were downloaded from and Megasense cloud and Prana website. Although the device makers calibrate these datasets, they are considered raw data in this study because the low-cost devices are calibrated in different environments and may not provide precise values; thus, the device makers suggest for the devices to be calibrated locally for accurate findings.

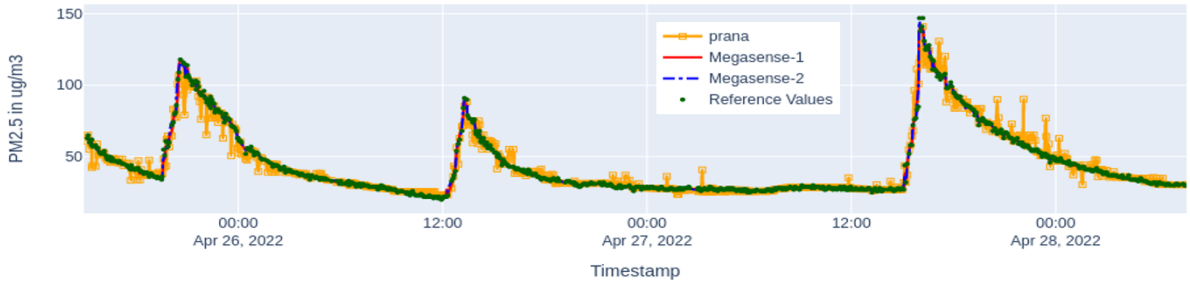
The time series plots for the raw and calibrated data from the three devices and the reference device for $PM_{2.5}$, CO, and NO_2 in the laboratory experiment are shown in Figs. 4.3, 4.4, 4.5 and mobile experiments are shown in Figs. 4.6, 4.7, 4.8. The Figs. 4.3(a), 4.4(a), 4.5(a), show the raw data time series plots in laboratory experiments; from the figures, it can be observed that the low-cost sensor types follow a similar trend as the reference device for all three pollutants. The maximum cross-correlation values of the MegaSense-1,2 at the lags of 7 min and 1 min, respectively, and Prana (no lag was observed) with the reference device for $PM_{2.5}$ are 0.98, 0.98, and 0.94. For CO, the maximum cross-correlation values of MegaSense-1,2 and Prana devices with the reference are 0.937, 0.91, and 0.973, respectively, at the lags of 6 min, 1 min, and 10 min, respectively. For NO_2 the maximum cross-correlation for the MegaSense-1, 2 (observed at no lag) and Prana (observed at 5 min lag) with the reference device are 0.29, 0.23, and 0.243, respectively.

Although the trends for the low-cost devices are similar to that of the reference device, there is a huge bias for MegaSense-1,2 as the values are in a different order. To visualize them in the same plot as the reference device, the values of MegaSense-1,2 are divided by 50 for CO, while for NO_2 they are divided by 10 and 50, respectively (these values were chosen through trial and error). This results in huge RMSE and MAE as seen from Table 4.2 where we show MegaSense-1 for ease of comparison against Prana, which has values in the same range and significant error values. Therefore, the Figs. 4.3(a), 4.4(a), 4.5(a) and Table 4.2 demonstrate the need for calibration of the low-cost sensor types. Moreover, it can be seen from the plots that each device (although it may be from the same manufacturer) has to be calibrated separately. Similar observations can be made about the mobile measurements from Table 4.2 and Figs. 4.6(a), 4.7(a), 4.8(a).

[2]Note that the range of raw CO and NO_2 values for MegaSense-1,2 are in a different order as compared to the reference device. Therefore, the raw CO values from MegaSense-1, 2 are divided by 50 while raw NO_2 values from MegaSense-1 & MegaSense-2 are divided by 10 & 50, respectively, so that all the trends can be easily visualized in the same plot.



(a) $PM_{2.5}$ raw data time-series plot.

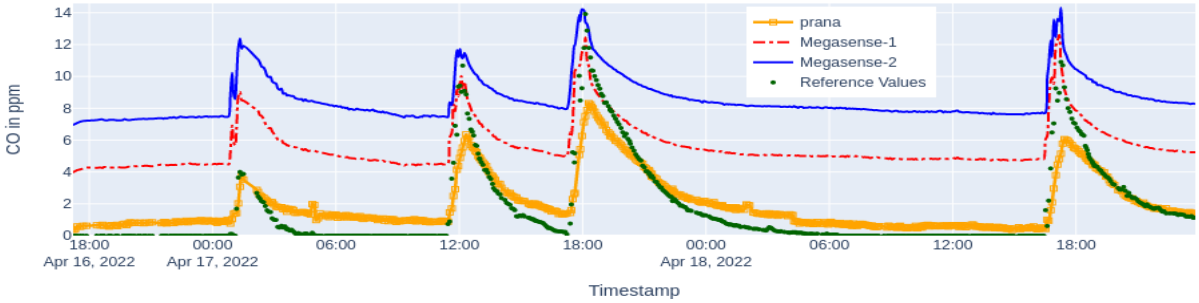


(b) $PM_{2.5}$ calibrated data time-series plot.

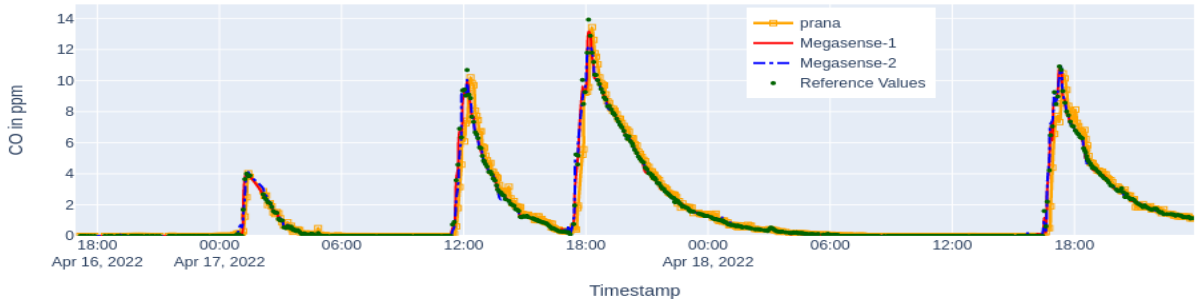
Figure 4.3: Raw and calibrated data time-series plots of $PM_{2.5}$ for all the 3 devices along with the reference device in laboratory settings[1].

Table 4.3: Cross validation performance metrics for PM indoor calibration

Sensor Type	Algorithm	Features				Performance Parameters		
		$PM_{2.5}$	Temp	RH	Pressure	R^2	MAE	RMSE
MegaSense - 1	LR	X				0.968	2.185	4.132
	MLR	X	X	X		0.969	2.11	4.06
	PR-7	X		X	X	0.978	1.69	3.38
	DT	X	X	X	X	0.974	1.624	3.612
	RF	**X**	**X**	**X**	**X**	**0.985**	**1.28**	**2.74**
	SVR	X	X	X	X	0.967	1.59	4.149
Prana	LR	X				0.885	5.343	7.831
	MLR	X	X	X		0.886	5.331	7.805
	PR - 5	X	X	X		0.927	3.79	6.22
	DT	X	X	X		0.918	3.08	6.621
	RF	**X**	**X**	**X**		**0.932**	**3.108**	**5.97**
	SVR	X	X	X		0.916	3.814	6.664



(a) CO raw data time-series plot.

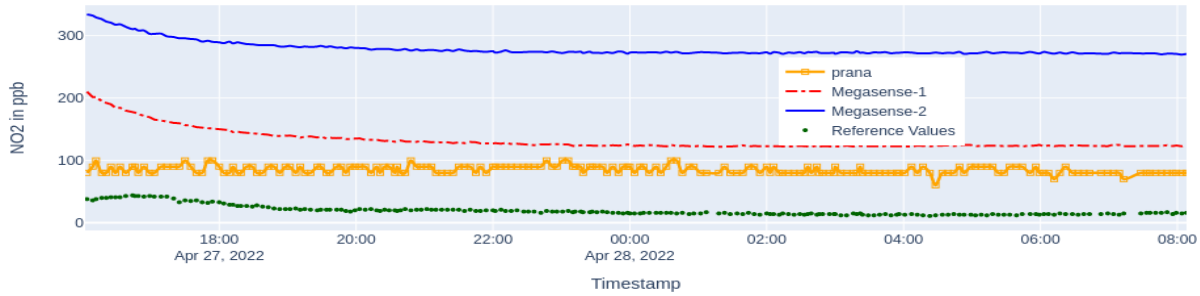


(b) CO calibrated data time-series plot.

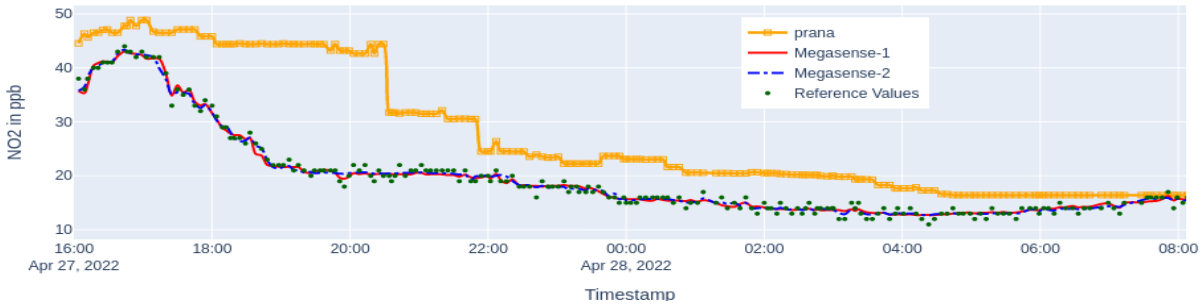
Figure 4.4: Raw and calibrated data time-series plots of CO for all the 3 devices along with the reference device in laboratory settings[2].

Table 4.4: Comparison of cross-validation performance metrics for CO laboratory calibration.

Sensor Type	Algorithm	Features				Performance Parameters		
		CO	Temp	RH	Pressure	R^2	MAE	RMSE
MegaSense-1	SLR	X				0.877	0.57	0.853
	MLR	X	X	X	X	0.914	0.473	0.702
	PR-13	X	X	X	X	0.98	0.165	0.34
	DT	X	X	X	X	0.975	0.103	0.376
	RFR-9	**X**	**X**	**X**	**X**	**0.987**	**0.08**	**0.28**
	SVR	X	X	X	X	0.927	0.22	0.65
Prana	SLR	X				0.947	0.45	0.608
	MLR	X	X	X		0.959	0.347	0.54
	PR-6	X	X	X		0.985	0.153	0.314
	DT	X	X	X		0.983	0.11	0.343
	RFR-9	**X**	**X**	**X**		**0.985**	**0.138**	**0.31**
	SVR	X	X	X		0.985	0.105	0.29



(a) NO_2 raw data time-series plot.



(b) NO_2 calibrated data time-series plot.

Figure 4.5: Raw and calibrated data time-series plots of NO_2 for all the three devices along with the reference device in laboratory settings.

Table 4.5: Comparison of cross validation performance metrics for NO_2 laboratory calibration.

Sensor Type	Algorithm	Features				Performance Parameters		
		NO_2	Temp	RH	Pressure	R^2	MAE	RMSE
MegaSense-1	SLR	X				0.179	9.73	11.52
	MLR	X	X	X	X	0.44	7.48	9.45
	PR-6	X	X	X		0.894	2.99	4.13
	DT	X	X	X	X	0.96	1.67	2.52
	RFR-9	**X**	**X**	**X**	**X**	**0.973**	**1.44**	**2.08**
	SVR	X	X	X	X	0.9	2.50	4.01
Prana	SLR	X				0.01	11.20	12.93
	MLR	X	X	X		0.646	6.27	7.68
	PR-5	X	X	X		0.904	2.87	3.99
	DT	X	X	X		0.922	2.01	3.48
	RFR-11	**X**	**X**	**X**		**0.958**	**1.53**	**2.62**
	SVR	X	X	X		0.946	1.64	2.96

4.4.2 Calibration Results

4.4.2.1 Laboratory Calibration

The calibration performance of different ML algorithms using the data collected for $PM_{2.5}$, CO, and NO_2 gases in the laboratory are depicted in Table 4.3, Table 4.4 and Table 4.5. For ML algorithms, different combinations of input features were considered. The selected features for the model are marked 'X' in the tables. The degree of the polynomial used in PR is indicated in the table for each device. For example, PR-3 is a third-degree polynomial. For RFR, the tree depth is chosen in the range of 2-20 such that R^2 score is maximum. For each algorithm, the results are presented only for the best combinations of input features for brevity.

The laboratory-calibrated data time series plots for $PM_{2.5}$, CO, and NO_2 are shown in Figs. 4.3(b), 4.4(b), and 4.5(b) respectively. It was observed that all sensor devices produce data very close to the reference instrument's following calibration. There is a significant improvement in the performance parameters after calibration compared to using raw observations (as shown in Table 4.2) for all three pollutants. From the Tables 4.3, 4.4, and 4.5, it can be observed that the ML algorithms PR, DT, RFR, and SVR give better performance compared to the linear algorithms (LR and MLR). Among the ML algorithms, RFR gives the best performance for all three pollutants.

MegaSense-1 device performs the best for all the three pollutants $PM_{2.5}$, CO, and NO_2. For $PM_{2.5}$, it gives the least cross-validation MAE and RMSE, of 1.28 ppm and 2.74 ppm, respectively, and the highest R^2 of 0.985. From Fig. 4.3(b), it can be observed that bias for all the devices is reduced. For CO, the least cross-validation MAE of 0.08 ppm, RMSE of 0.28 ppm, and highest R^2 of 0.987. From Fig. 4.4(b), it can be observed that the bias for all the devices is reduced, and the readings of the devices are in the same range as that of the reference device. The sudden increase in CO values is observed when the incense stick is lit, and the values decrease when the smoke is allowed to leak out of the controlled environment. For NO_2, the least cross-validation MAE of 1.44 ppb, RMSE of 2.08 ppb, and the highest R^2 of 0.973. All the sensor devices after calibration have R^2 greater than 0.94 and MAE and RMSE less than 1.91 ppb and 2.94 ppb, respectively. These values show a great improvement when compared to the raw values' performance from Table 4.2. Similar to CO from Fig. 4.5(b), it can be observed that the bias for all the devices is reduced, and the readings of the devices are in the same range as that of the reference device.

4.4.2.2 Application of Laboratory ML Models on Mobile Data

The performance of the devices when the laboratory-calibrated model is applied to the mobile test data is shown in Table 4.6. It can be observed that the error is lesser compared to the error using the raw mobile measurements as shown in Table 4.2 for all the pollutants $PM_{2.5}$, CO, and NO_2. However,

these errors can be further reduced when the devices are calibrated using the mobile ML models on the mobile data, as will be demonstrated next.

Table 4.6: Comparison of performance for laboratory calibration applied on raw mobile test data.

	$PM_{2.5}$		CO		NO_2	
Sensor Type	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
MegaSense-1	22.09	30.99	6.08	6.84	33.74	41.77
Prana	36.94	48.59	4.58	5.69	50.17	64.39

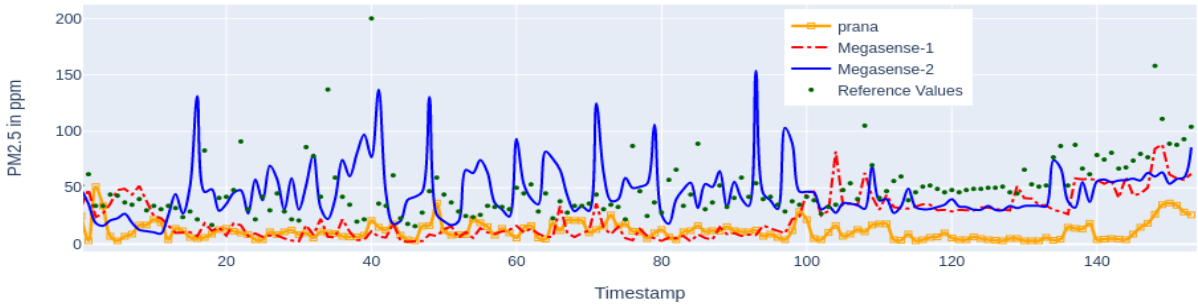
4.4.2.3 Mobile Calibration

The calibration performance of different ML algorithms for $PM_{2.5}$, CO, and NO_2, respectively, while using the mobile data for training instead of laboratory data are shown in Tables 4.7, 4.8 and 4.9. Three important observations can be made. First, there is a significant decrease (50-75 %) in MAE and RMSE when we use mobile ML models as compared to laboratory ML models for the mobile data. Second, among the ML algorithms considered, RFR performs the best for both gases while linear algorithms perform the worst. Although the performance of calibration algorithms for CO is similar in terms of MAE and RMSE, the performance is very poor except for RFR in terms of R^2. For NO_2 the performance of RFR is better in terms of all three parameters significantly. Among the three devices, MegaSense-1 performs the best while Prana's performance is unsatisfactory.

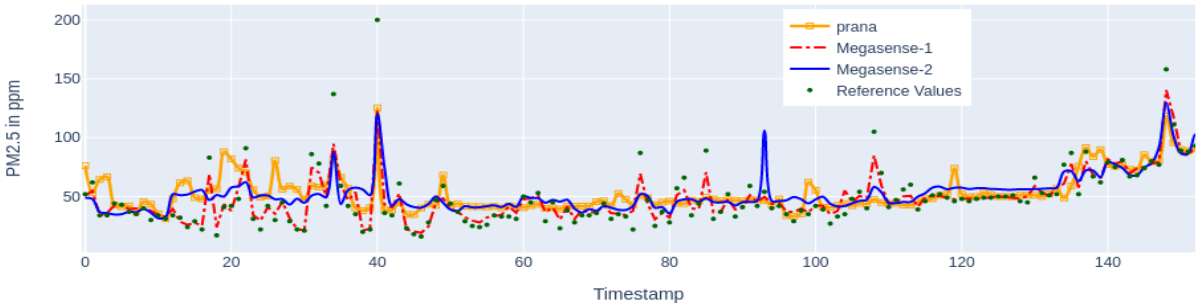
It can be observed that MegaSense-1 performs the best among all the three devices for all the three pollutants. When the devices are calibrated using SLR, it is seen that the error has reduced significantly for MegaSense-1,2; whereas there is not much difference for Prana. Each device needs to be calibrated separately is necessary to ensure accuracy and consistency in measurements or performance. Each device, even if it's the same model from the same manufacturer, may have slight variations in its components or manufacturing process. When additional features like temperature, RH, and pressure are added to train the model the error MAE, RMSE for MegaSense-1,2 is further reduced and slight improvement in R^2 can be observed, for Prana no significant change observed in terms of error. The error is further reduced and the R^2 is further improved in case of RFR for MegaSense-1,2. Even though R^2 score is improved for Prana it is not satisfactory, and the error has barely changed. RFR performs the best for all the three devices.

By comparing the mobile performance of the devices in Table 4.6 and data in Tables 4.7, 4.8 & 4.9 it can be said that the performance of the devices has been improved in terms of the error metrics for all the three pollutants . In the case of $PM_{2.5}$, error for MegaSense -1 has nearly decreased by two times, for prana the error has reduced nearly by 0.6 times. The error has almost decreased by three times for all the devices in the case of CO. For NO_2 the error has decreased when compared to Table 4.6.

The line plots for raw and calibrated data for mobile measurements for $PM_{2.5}$, CO, and $NO_2$are shown in Figs. 4.6, 4.7 and 4.8 respectively. In the case of raw mobile data, it can be observed that all three devices did not follow the same trend as that of the reference device. Once calibrated, all



(a) $PM_{2.5}$ raw data time-series plot.

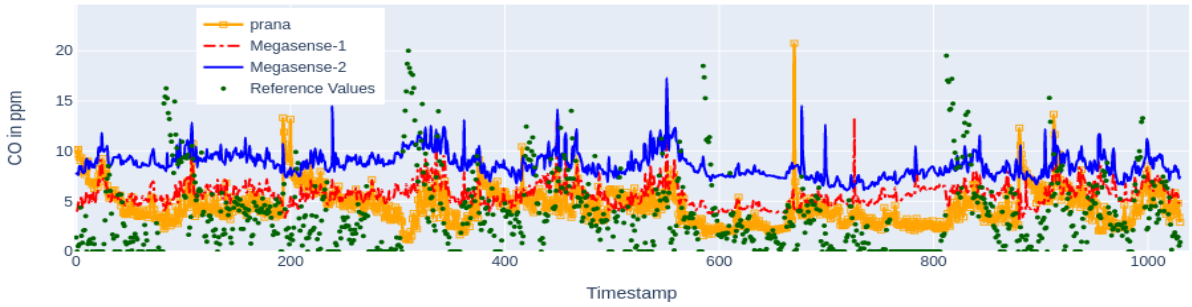


(b) $PM_{2.5}$ calibrated data time-series plot.

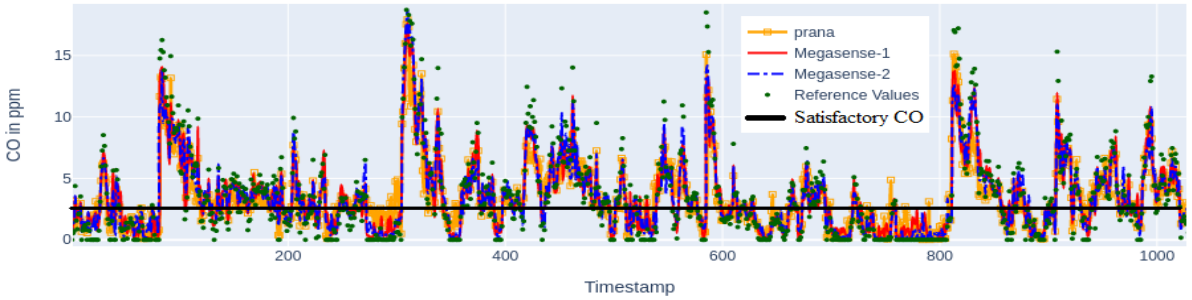
Figure 4.6: Raw and calibrated data time-series plots of $PM_{2.5}$ for all the 3 devices along with the reference device in mobile settings.

Table 4.7: Comparison of cross-validation performance metrics for $PM_{2.5}$ mobile calibration.

Sensor Type	Algorithm	Features				Performance Parameters		
		$PM_{2.5}$	Temp	RH	Pressure	R^2	MAE	RMSE
MegaSense-1	SLR	X				0.279	14.94	20.92
	MLR	X	X	X	X	0.437	13.07	18.45
	PR-3	X			X	0.468	11.81	17.96
	DT	X	X	X	X	0.306	12.66	20.52
	RFR-12	**X**	**X**	**X**	**X**	**0.62**	**9.71**	**15.19**
	SVR	X	X	X	X	0.344	13.26	20.19
Prana	SLR	X				0.072	26.40	40.07
	MLR	X	X	X		0.167	24.32	38.08
	PR-3	X	X			0.17	24.78	37.18
	DT	X				-0.227	28.65	44.52
	RFR-4	**X**	**X**	**X**		**0.286**	**20.33**	**34.08**
	SVR	X	X	X		0.271	20.87	34.60



(a) CO raw data time-series plot.

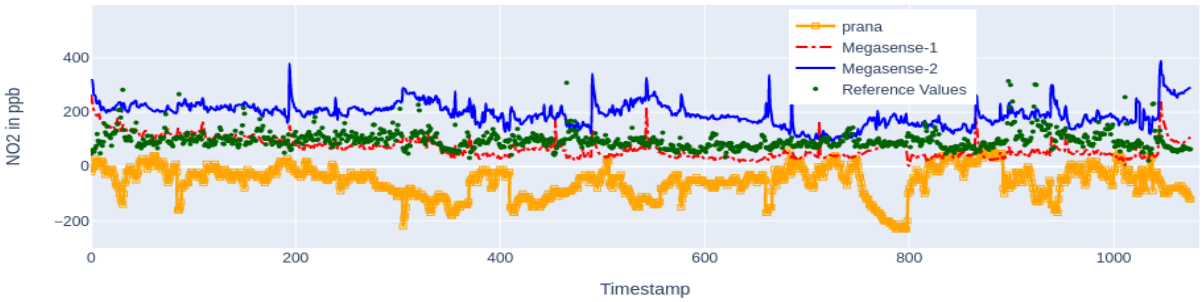


(b) CO calibrated data time-series plot.

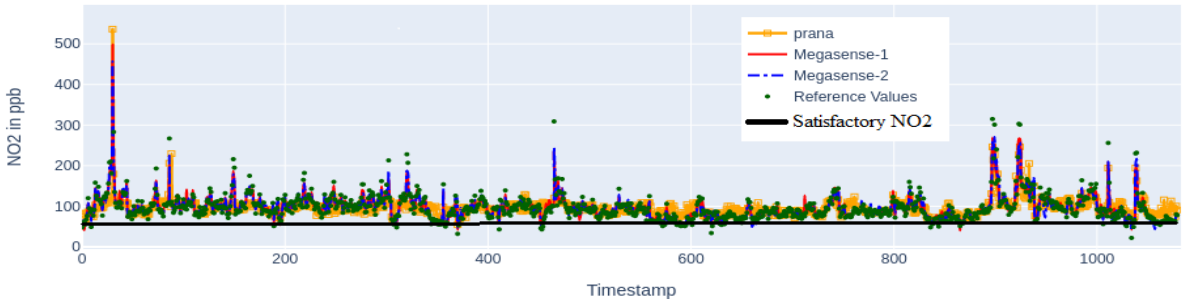
Figure 4.7: Raw and calibrated data time-series plots of CO for all the 3 devices along with the reference device in mobile settings[2].

Table 4.8: Comparison of cross-validation performance metrics for CO mobile calibration.

Sensor Type	Algorithm	Features				Performance Parameters		
		CO	Temp	RH	Pressure	R^2	MAE	RMSE
MegaSense-1	SLR	X				0.202	2.62	3.70
	MLR	X	X	X	X	0.22	2.55	3.66
	PR-4	X			X	0.258	2.42	3.56
	DT	X	X	X	X	0.337	1.82	3.34
	RFR-16	**X**	**X**	**X**	**X**	**0.66**	**1.44**	**2.41**
	SVR	X	X	X	X	0.249	2.28	3.59
Prana	SLR	X				-0.009	2.84	3.76
	MLR	X	X	X		0.048	2.77	3.65
	PR-3	X	X	X		0.076	2.66	3.60
	DT	X	X	X		-0.145	2.53	3.95
	RFR-11	**X**	**X**	**X**		**0.353**	**2.09**	**3.01**
	SVR	X	X	X		0.181	2.37	3.35



(a) NO_2 raw data time-series plot.

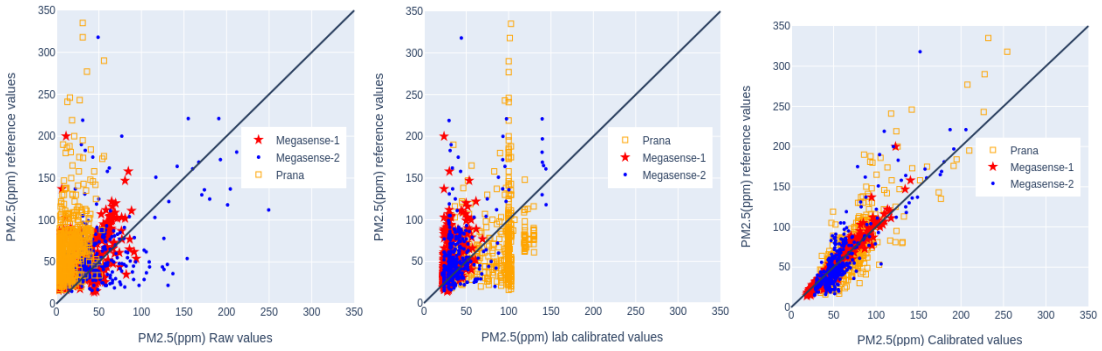


(b) NO_2 calibrated data time-series plot.

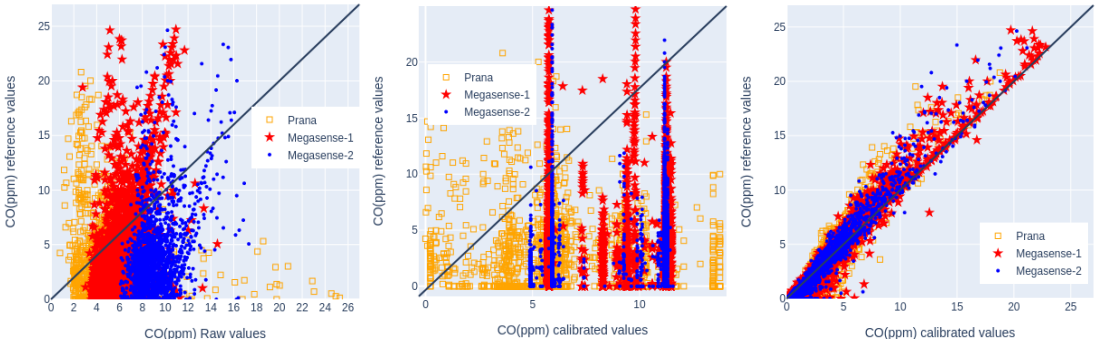
Figure 4.8: Raw and calibrated data time-series plots of NO_2 for all the 3 devices along with the reference device in mobile settings.

Table 4.9: Comparison of cross validation performance metrics for NO_2 mobile calibration.

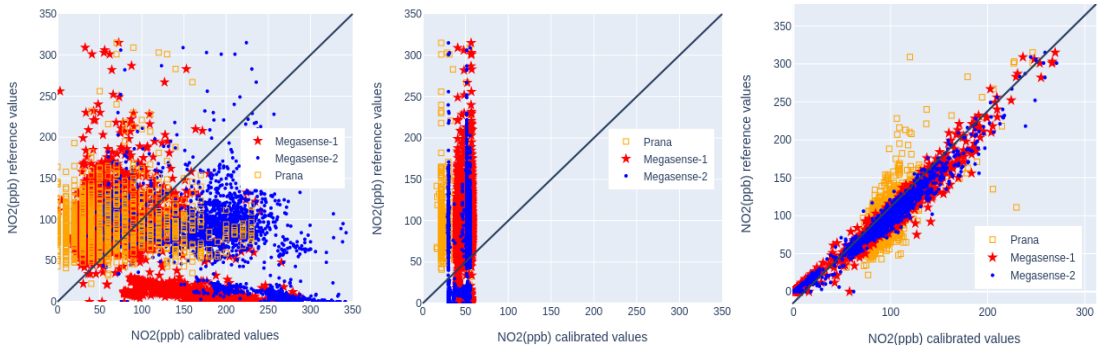
Sensor Type	Algorithm	Features				Performance Parameters		
		NO_2	Temp	RH	Pressure	R^2	MAE	RMSE
MegaSense-1	SLR	X				0.365	30.77	41.95
	MLR	X	X	X	X	0.573	23.37	34.45
	PR-4	X		X	X	0.658	18.55	30.43
	DT	X	X	X	X	0.672	15.53	30.07
	RFR-20	**X**	**X**	**X**	**X**	**0.813**	**11.60**	**22.73**
	SVR	X	X	X	X	0.372	28.22	41.70
Prana	SLR	X				-0.004	23.03	36.19
	MLR	X	X	X		0.014	22.95	35.91
	PR-10	X	X	X		0.095	22.20	34.18
	DT	X				0.019	22.63	35.79
	RFR-8	**X**	**X**	**X**		**0.105**	**21.32**	**33.81**
	SVR	X	X	X		0.069	21.30	34.08



(a) $PM_{2.5}$ mobile raw data scatter plot. (b) $PM_{2.5}$ lab-model on mobile data scatter plot. (c) $PM_{2.5}$ mobile calibrated data scatter plot.



(d) CO mobile raw data scatter plot. (e) CO lab-model on mobile data scatter plot. (f) CO mobile calibrated data scatter plot.



(g) NO_2 mobile raw data scatter plot. (h) NO_2 lab-model on mobile data scatter plot. (i) NO_2 mobile calibrated data scatter plot.

Figure 4.9: $PM_{2.5}$, CO, and NO_2 raw and calibrated data scatter plots for mobile experiments.

the devices followed a similar trend as that of the reference device, and the bias was also reduced significantly. The $PM_{2.5}$, CO, and NO_2 values for 24 hr, 1 hr, and 24 hr average according to Indian standards [83] should be less than 30 ppm, 3.4 ppm, and 42.55 ppb respectively, and are shown in Figs. 4.6(b), 4.7(b), and 4.8(b) respectively . From the plots, it can be observed that the majority of the $PM_{2.5}$ and CO values are above the standard values, while for NO_2, almost all the values are above the standard level.

Fig. 4.9 shows the scatterplot for $PM_{2.5}$, CO and NO_2 mobile data for the three cases: raw data, calibrated data using lab models, and calibrated data using mobile models. The best-performing model results are used to plot the calibrated data scatter plot. From the Figs. 4.9(a),4.9(d) and 4.9(g), it can be said that all the devices deviate from the 1:1 line for raw data. Figs. 4.9(b), 4.9(e) and 4.9(h) show the scatterplot of the data when the lab calibrated models are applied to the mobile data for $PM_{2.5}$, CO and NO_2 respectively, it can be observed that data is clustered at certain range of values and do not follow linearity with the reference in all the three cases. Figs. 4.9(c), 4.9(f) and 4.9(i) show scatterplots of the devices' mobile calibrated data. Once calibrated, the MegaSense devices follow linearity and are in the same range as the reference device for $PM_{2.5}$, CO and NO_2, while Prana continues to deviate from the 1:1 line in the case of NO_2. Also, it can be seen that after calibration the Megasense devices are in the same range as that of the reference device.

4.4.3 Outcome of the Analysis

4.4.3.1 Diwali Data Analysis

The mobile data from all low-cost devices were collected in various parts of the city during the festival of Diwali from November 1^{st} to 7^{th}. The mobile calibrated functions were applied to the raw values of the mobile measurement campaign, and the obtained calibrated values were used for the analysis. During this week, firecrackers are commonly set off in almost every part of the city. As a result, the range of pollutant concentrations in the air is higher than usual. The pollutant concentrations from mobile data obtained this week for both contaminants are plotted on the map for MegaSense devices as they are performing the best. $PM_{2.5}$ concentration shown on the map in Fig:4.10(a). PM_{10} and $PM_{2.5}$ levels are classified into five categories based on the first five value ranges stated in[84]: Good, Satisfactory, Moderate, Poor, and Very Poor. $PM_{2.5}$ values are in all ranges of the AQI categories. The $PM_{2.5}$ levels are in the range 22- 145 ug/m3. The PM values are observed to be high during the night time during the festival days, when there were muddy roads, high traffic areas. High levels of $PM_{2.5}$ were detected at night on the third and fourth days of the measurement campaign, when fireworks were lighted to commemorate the festival in the city, as well as in regions with heavy traffic and congested surroundings. At-risk persons should avoid all outdoor activities. All others should avoid prolonged exertion outdoors. This includes the drivers and passengers in the vehicles. Higher $PM_{2.5}$ concentrations were identified along the ring expressway and points due to traffic congestion. Values between 55 to 150 are

unhealthy for everyone leading to respiratory aggravation in the general population. At-risk children and the elderly should avoid exertion. Everyone else should limit exertion.

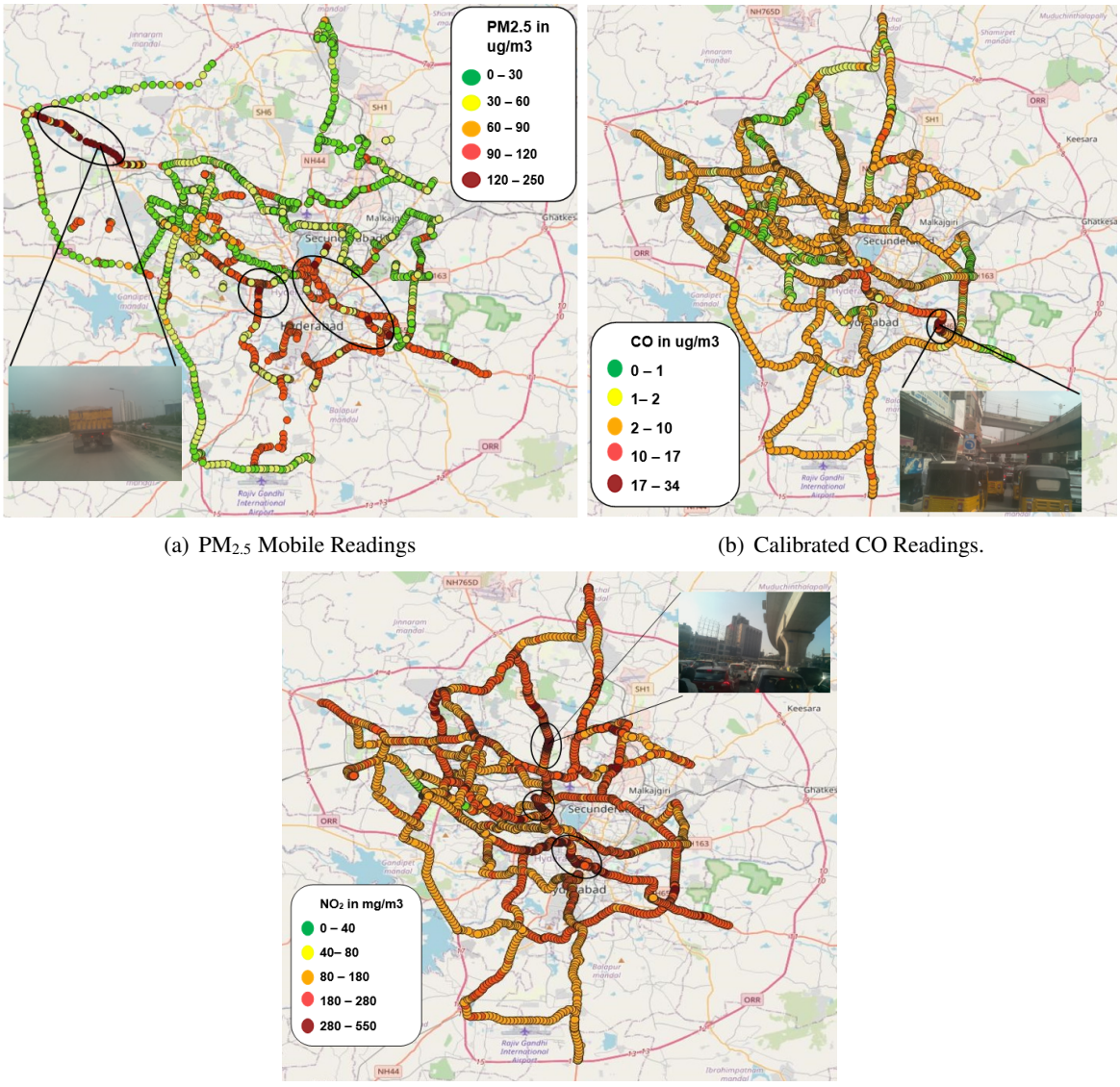
Fig:4.10(b) shows the CO values along the path traveled during the measurement campaign. The CO readings are divided into 5 categories based on the [84]. The majority of the readings, as seen in the graph, fall into the moderate CO levels. CO levels were discovered to be high at night at about 9 p.m. on the 3^{rd} and 4^{th} of November, which was on the festival day when firecrackers were lit. The CO values detected are in the range of 0 to 19 ppm. CO levels were found to be low near highways, locations with little or no traffic, and areas with very low human density. Moderate CO levels are seen in regions with less traffic and population density, such as village areas and open spaces. High values of CO were observed during the nighttime on the third, and fourth day of the measurement campaign when the fireworks were lit to celebrate the festival in the city, as well as in the areas where there was high traffic with congested surroundings.

The NO_2 values along the path traveled during the measurement campaign are depicted in Fig. 4.10(c). The NO_2 levels detected range from 0 to 135 ppb. Based on the [84], the NO_2 readings are classified into five categories. From the 4.10(c), it can be observed that almost all the readings fall into poor and very poor groups of NO_2 levels. High values of NO_2 were observed when the traffic was high, near the industrial regions. The advice for people in the streets and living beside busy roads should be to consider limiting prolonged outdoor exertion.

4.4.3.2 CO and NO_2 Emission Spike Detection

The mobile sensing data sets used to calibrate the IoT sensors were reevaluated in order to detect CO and NO_2 emission hotspots to be located at a high resolution in the city of Hyderabad and recommendations for local people to be shared.

Spatial distribution of emission spikes: The street vehicle contributed to the traffic conditions. The onboard mobile sensing data indicates that road network characteristics: road type and the number of vehicles the road can handle and connection points such as junctions and traffic lights - are important ephemeral precursors to CO and NO_2 emission spikes. Just as important are urban Points of Interest (POIs), which people visit for short periods, such as hospitals, schools, and temples. These POIs are often characterized by different vehicle categories, erratically stopping and leaving, traveling at different speeds, and temporally parking beside the road, blocking traffic flow and creating CO and NO_2 emission spikes. The mobile sensing in Hyderabad identified emission spikes in Hyderabad where there was high building density, at road network junctions, and at POIs depicted in Fig. 4.11, and 4.12. Data collected on January 4th show high NO_2 measurements at POI (temple, garden, and parks) near Hussain Sagar Lake. Inferring high visitor stop-go traffic behavior disrupting regular traffic flow. Data collected on January 6th (map not shown) detected a CO emission spike near the international airport. And CO and NO_2 emission spikes on the primary roads connected the outer ring road (ORR), an 8-lane ring outer ring expressway. Data collected on January 8th: CO and NO_2 emission spikes in both densely built areas near the temple and more open areas with high values near hospitals and medical centers.

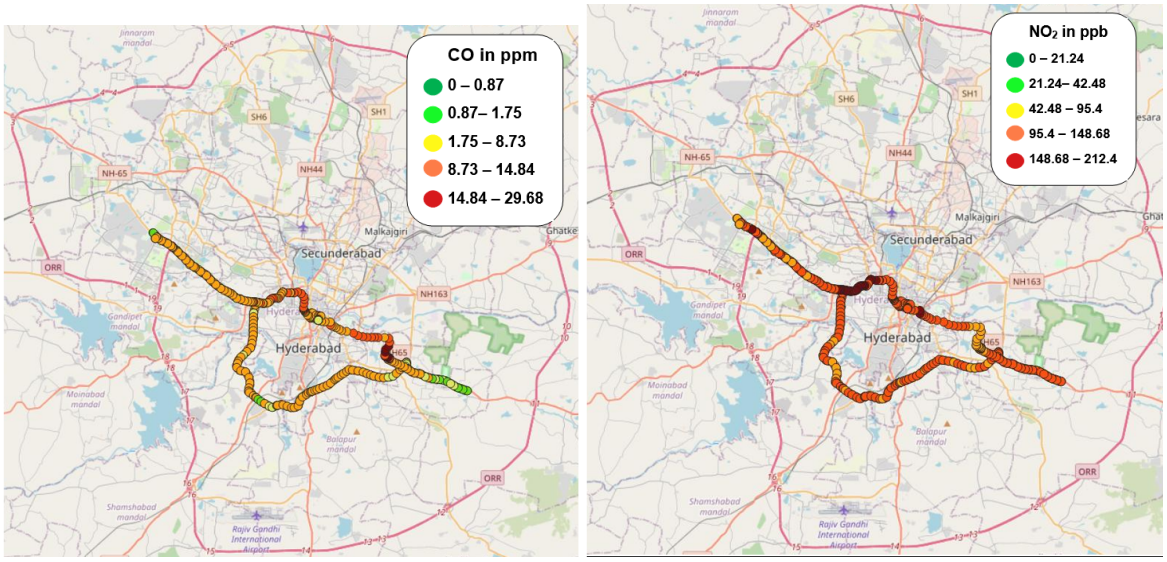


(a) $PM_{2.5}$ Mobile Readings

(b) Calibrated CO Readings.

(c) Calibrated NO_2 Readings.

Figure 4.10: Mobile measurements of CO and NO_2 during the festival of Diwali, 2021.

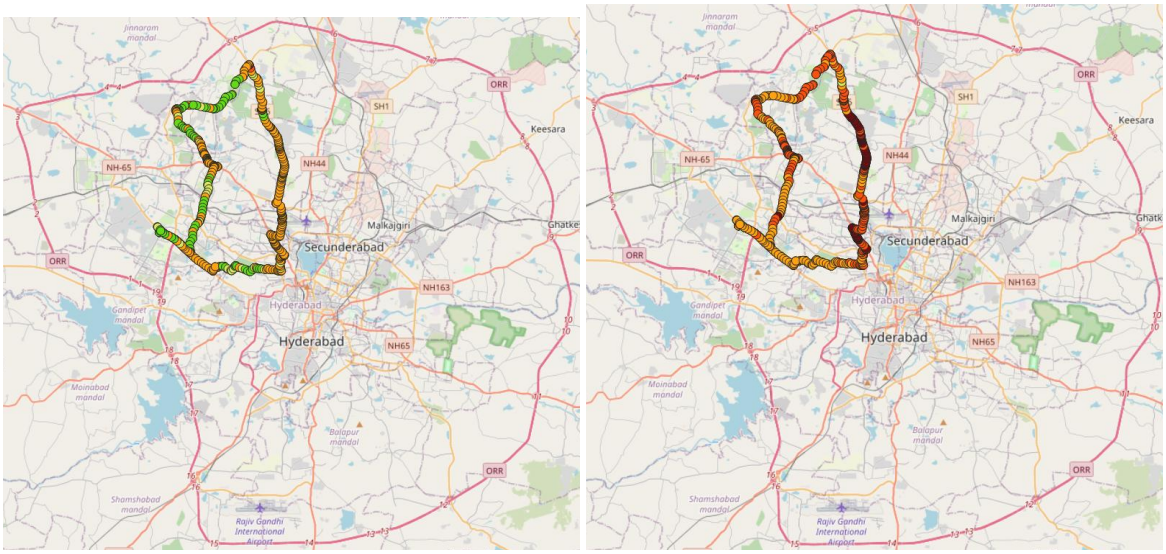


(a) 4 Jan CO

(b) 4 Jan NO_2

Figure 4.11: CO & NO_2 mobile sensing measurements and location of emission spikes on 4th January

Temporal variability of emission spikes:The mobile sensing data captured the temporal variability of air pollution in Hyderabad. Making daily CO and NO_2 city measurements highlights the importance of meteorological conditions where one day the air pollution is persistently high and the next day the values are persistently moderate, even the daily urban patterns are quite regular. Driving throughout the city, high levels of CO does not always correspond with high levels of NO_2 although there is a causal relationship between CO and NO_2 because CO slowly oxidizes NO to NO_2. CO is also emitted by households and burning trash, whereas NO_2 is predominately emitted by heavy traffic. The temporal analysis can be used to inform citizens when they should avoid high levels of unhealthy exposure to air pollutants. For example, data collected on January 4th detected persistent air pollution periods when CO peaked above 10 ppm and NO_2 ranging between 70 to 160 ppb, both exceeding the EU and WHO limit values (Fig. 4.5(b)) (same calibration methods are applied to both Diwali and January data). During the bad periods when NO_2 gas concentrations were between 101-150 ppb, people with lung disease and children and older adults exposed should have limited prolonged outdoor exertion. Whereas, during unhealthy periods when NO_2 concentrations exceeded 150 (150- 200 ppb), people with lung disease and children and older adults should have avoided prolonged outdoor exertion. On January 8th daily CO concentrations were acceptably below 5 ppm, but NO_2 was unhealthily high, ranging between 50 to 190 ppb. The advice for people in the streets and living beside busy roads should be to consider limiting prolonged outdoor exertion.



(a) 8 Jan CO

(b) 8 Jan NO_2

Figure 4.12: CO & NO_2 mobile measurements and location of emission spikes on 8^{th} January

Chapter 5

Concluding Remarks

This thesis underscores the importance of calibrating low-cost air pollution monitoring IoT devices using ML and integrating mobile measurements to assess air quality and identify pollutant hotspots. It elucidates the pivotal role of calibration in ensuring precise and dependable readings. The research establishes that lacking proper calibration can introduce significant biases and inaccuracies into data from low-cost sensors, ultimately leading to erroneous conclusions. Consequently, the calibration process emerges as a crucial step in harnessing the full potential of these sensors.

Moreover, the significance of mobile measurements in understanding air pollution patterns is emphasized within this thesis. By leveraging monitoring instruments' mobility and adaptability, real-time air quality disparities across diverse locations can be effectively collected. This capability enables pinpointing specific areas or hotspots.

The study showcases the potential of IoT devices integrated into vehicles for gathering air quality data. It offers a framework for identifying transient and persistent emission spikes in polluted urban environments by evaluating IoT devices mounted on a mobile platform, such as a streetcar, in Hyderabad. The research indicates that Random Forest Regression (RFR) calibration yields the best results due to the high variability in mobile sensing datasets. Additionally, it reveals that points of interest causing traffic disruptions play a significant role in generating spikes of unhealthy $PM_{2.5}$, CO, and NO_2 concentrations.

Nevertheless, the study acknowledges certain limitations. The laboratory calibration methodology could have employed a more scientifically rigorous approach rather than relying on an incense stick for NO_2 emission. Further enhancement could involve incorporating more IoT devices and conducting multiple streetcars runs on different days to yield higher-quality data. A more robust experimental design and a concentrated focus on monitoring air pollution and taking action at points of interest within polluted urban areas are recommended for future research.

Related Publications

Conference Papers:

- Spanddhana Sara, Andrew Rebeiro-Hargrave, Shreyash Gujar, Om Kathalkar, Samu Varjonen, Sachin Chaudhari, and Sasu Tarkoma. 2023. "Protocol for hunting PM2.5 emission hot spots in cities". In 1st International Workshop on Advances in Environmental Sensing Systems for Smart Cities (EnvSys '23), June 18, 2023, Helsinki, Finland. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3597064.3597322
- Ishan Patwardhan, Spanddhana Sara, Sachin Chaudhari. "Comparative Evaluation of New Low-Cost Particulate Matter Sensors," 2021 8th International Conference on Future Internet of Things and Cloud (FiCloud), Rome, Italy, 2021, pp. 192-197, doi: 10.1109/FiCloud49777.2021.00035.

Journals:

- Spanddhana Sara, Andrew Rebeiro-Hargrave, Ayu Parmar, Pak Lun Fung, Ishan Patwardhan, Samu Varjonen, C. Rajashekar Reddy, Sachin Chaudhari, and Sasu Tarkoma. "The Application of Mobile Sensing to Detect CO and NO_2 Emission Spikes in Polluted Cities," in IEEE Access, vol. 11, pp. 79624-79635, 2023, doi: 10.1109/ACCESS.2023.3297874.

Other Publications:

- Ayu Parmar, Spanddhana Sara, Ayush Kumar Dwivedi, C. Rajashekar Reddy, Ishan Patwardhan, Sai Dinesh Bijjam, Sachin Chaudhari, K. S. Rajan, Kavita Vemuri. "Development of end-to-end low-cost IoT system for densely deployed PM monitoring network: an Indian case study, " in Frontiers Internet Things. doi: 10.3389/friot.2024.1332322.

Bibliography

[1] "Around 3 billion people cook and heat their homes using polluting fuels," accessed 9 Aug, 2022, https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/health-impacts/types-of-pollutants.

[2] D. Saadi, E. Tirosh, and I. Schnell, "The relationship between city size and carbon monoxide (CO) concentration and their effect on heart rate variability (hrv)," *International Journal of Environmental Research and Public Health*, vol. 18, no. 2, p. 788, 2021.

[3] T. V. Kokkonen et al, "The effect of urban morphological characteristics on the spatial variation of PM 2.5 air quality in downtown nanjing," *Environmental Science: Atmospheres*, vol. 1, no. 7, pp. 481–497, 2021.

[4] J. Bi et al, "Within-city variation in ambient carbon monoxide concentrations: Leveraging low-cost monitors in a spatiotemporal modeling framework," vol. 130, no. 9, p. 097008, 2022.

[5] N. H. Motlagh et al, "Toward massive scale air quality monitoring," *IEEE Communications Magazine*, 2020.

[6] A. Gonzalez, A. Boies, J. Swason, and D. Kittelson, "Field calibration of low-cost air pollution sensors," *Atmospheric Measurement Techniques Discussions*, 2019.

[7] J. S. Apte et al., "High-resolution air pollution mapping with google street view cars: Exploiting big data," *Environmental Science & Technology*, 2017.

[8] A. Parmar et al., "Development of end-to-end low-cost iot system for densely deployed pm monitoring network: An indian case study," *Frontiers Journal, IoT Services and Applications*. [Online]. Available: https://doi.org/10.3389/friot.2024.1332322

[9] M. I. Mead et al., "The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks," *Atmospheric Environment*, 2013.

[10] K. Sadighi et al., "Intra-urban spatial variability of surface ozone in riverside, CA: viability and validation of low-cost sensors," *Atmospheric Measurement Techniques*, 2018.

[11] A. R. Al-Ali, I. Zualkernan, and F. Aloul, "A mobile gprs-sensors array for air pollution monitoring," *IEEE Sensors Journal*, 2010.

[12] M. D. Adams and D. Corr, "A mobile air pollution monitoring data set," *Data*, 2019.

[13] N. H. Motlagh et al, "Transit pollution exposure monitoring using low-cost wearable sensors," *Transportation Research Part D: Transport and Environment*, vol. 98, p. 102981, 2021.

[14] A. Rebeiro-Hargrave et al, "City wide participatory sensing of air quality," *Frontiers in Environmental Science*, p. 587, 2021.

[15] F. Concas et al, "Low-cost outdoor air quality monitoring and sensor calibration: A survey and critical analysis," *ACM Transactions on Sensor Networks (TOSN)*, vol. 17, no. 2, pp. 1–44, 2021.

[16] N. H. Motlagh et al, "Low-cost air quality sensing process: Validation by indoor-outdoor measurements," in *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. IEEE, 2020, pp. 223–228.

[17] R. Kamal, *Internet of Things: Architecture and Design Principles*. Mc Graw Hill Education, 2017.

[18] P. Lea, *Internet of Things for Architects*. Birmingham, UK: Packt Publishing Ltd, 2018.

[19] C. C. Sobin, "A Survey on Architecture, Protocols and Challenges in IoT," *Wireless Personal Communications*, vol. 112, 2020.

[20] D. Norris, *The Internet of Things*. Mc Graw Hill Education, 2015.

[21] A. Bahgya and V. Madisetti, *Internet of Things: A Hands-on Approach*. Universities Press, 2015.

[22] TechTarget, https://www.techtarget.com/iotagenda/definition/Internet-of-Things-IoT.

[23] Olimex, https://www.olimex.com/Products/IoT/_images/thumbs/310x230/internet-of-things.jpg.

[24] Educba, https://cdn.educba.com/academy/wp-content/uploads/2019/12/iot-features.png.

[25] J. Lin et al, "A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1125–1142, 2017.

[26] Google Nest Thermostat, https://store.google.com/us/product/nest_thermostat.

[27] Amazon Alexa, https://www.amazon.com/alexa-smart-home/b?ie=UTF8&node=21442899011.

[28] "The internet of things in healthcare: An overview," *Journal of Industrial Information Integration*, vol. 1, pp. 3–13, 2016.

[29] Apple Healthcare, https://www.apple.com/in/healthcare/apple-watch/.

[30] M. Dhanaraju et al, "Smart farming: Internet of things (iot)-based sustainable agriculture," *Agriculture*, vol. 12, no. 10, 2022. [Online]. Available: https://www.mdpi.com/2077-0472/12/10/1745

[31] Microsoft FarmBeats, https://www.microsoft.com/en-us/research/project/farmbeats-iot-agriculture/.

[32] Honeywell, https://sps.honeywell.com/us/en/products/automation.

[33] Siemens, https://resources.sw.siemens.com/en-US/e-book-insight-hubs-industrial-iot-solutions-with-automated-low linkId=300000002500668.

[34] M. Whaiduzzaman et al, "A review of emerging technologies for iot-based smart cities," *MDPI Sensors*, 2022.

[35] "Internet of things (iot) applications to fight against covid-19 pandemic," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 4, pp. 521–524, 2020.

[36] Smartcity Living Labs, https://smartcityresearch.iiit.ac.in.

[37] K. S. Viswanadh, O. Kathalkar, P. Vinzey, N. Nilesh, S. Chaudhari, and V. Choppella, "Cv and iot-based remote triggered labs: Use case of conservation of mechanical energy," in *2022 9th International Conference on Future Internet of Things and Cloud (FiCloud)*, 2022, pp. 100–106.

[38] A. K. Lall, A. Khandelwal, R. Bose, N. Bawankar, N. Nilesh, A. Dwivedi, and S. Chaudhari, "Making analog water meter smart using ml and iot-based low-cost retrofitting," in *2021 8th International Conference on Future Internet of Things and Cloud (FiCloud)*, 2021, pp. 157–162.

[39] G. Ihita, K. Viswanadh, Y. Sudhansh, S. Chaudhari, and S. Gaur, "Security analysis of large scale iot network for pollution monitoring in urban india," in *2021 IEEE 7th World Forum on Internet of Things (WF-IoT)*, 2021, pp. 283–288.

[40] T. Hastie et al, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[41] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, http://www.deeplearningbook.org.

[42] M. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects." *Science*, vol. 349, no. 6245, p. 255–260, jul 2015.

[43] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[44] https://wordstream-files-prod.s3.amazonaws.com/s3fs-public/machine-learning.png.

[45] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*. AMLBook, 2012.

[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," ser. NIPS'12. Red Hook, NY, USA: Curran Associates Inc., 2012, p. 1097–1105.

[47] A. D. Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," in *2015 IEEE Symposium Series on Computational Intelligence*, 2015, pp. 159–166.

[48] A. Rajkomar, J. Dean, and I. S. Kohane, "Machine learning in medicine," *The New England Journal of Medicine*, vol. 380, p. 1347–1358, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:92996321

[49] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[50] M. Bojarski et al, "End to end learning for self-driving cars." *CoRR*, vol. abs/1604.07316, 2016.

[51] "Measuring pollution: How government is ramping up air quality monitoring stations according to Indian standards," https://economictimes.indiatimes.com/news/politics-and-nation/measuring-pollution-how-government-is-ramping-up-air-quality -monitoring-stations-according-to-indian-standards/articleshow/72410188.cms.

[52] X. Xie and I. Semanjski et al, "A review of urban air pollution monitoring and exposure assessment methods," *ISPRS International Journal of Geo-Information*, vol. 6, no. 12, 2017. [Online]. Available: https://www.mdpi.com/2220-9964/6/12/389

[53] C. R. Reddy, T. Mukku, and A. Dwivedi et al, "Improving spatio-temporal understanding of particulate matter using low-cost iot sensors," in *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, 2020, pp. 1–7.

[54] London Air Quality Network, "London air quality network—summary report 2018," *Techinal Report Environmental Research Group, King's College London: London, UK*, 2018.

[55] R. N. Murty et al., "CitySense: An Urban-Scale Wireless Sensor Network and Testbed," in *IEEE Conference on Technologies for Homeland Security*, 2008, pp. 583–588.

[56] D. Hasenfratz, O. Saukh, S. Sturzenegger, and L. Thiele et al., "Participatory air pollution monitoring using smartphones," *Mobile Sensing*, vol. 1, pp. 1–5, 2012.

[57] S. Kumar and A. Jasuja, "Air quality monitoring system based on IoT using Raspberry Pi," in *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 2017, pp. 1341–1346.

[58] A. Anjomshoaa et al., "City Scanner: Building and Scheduling a Mobile Sensing Platform for Smart City Services," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 4567–4579, 2018.

[59] C. Yanju et al., "A new mobile monitoring approach to characterize community-scale air pollution patterns and identify local high pollution zones," *Atmospheric Environment*, 2022.

[60] J. Van den Bossche et al., "Mobile monitoring for mapping spatial variation in urban air quality: Development and validation of a methodology based on an extensive dataset," *Atmospheric Environment*, vol. 105, pp. 148–161, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1352231015000254

[61] I. Wadlow et al., "Understanding spatial variability of air quality in sydney: Part 2 – A Roadside Case Study," *Atmosphere*, vol. 10, no. 4, 2019. [Online]. Available: https://www.mdpi.com/2073-4433/10/4/217

[62] N. Nilesh, I. Patwardhan, J. Narang, and S. Chaudhari, "Iot-based aqi estimation using image processing and learning methods," in *2022 IEEE 8th World Forum on Internet of Things (WF-IoT)*, 2022, pp. 1–5.

[63] K. N Genikomsakis et al, "Development and on-field testing of low-cost portable system for monitoring pm2.5 concentrations," *MDPI Sensors*.

[64] Prana, https://www.pranaair.com/air-quality-sensor/outdoor-pm-sensor/.

[65] Sensirion, https://www.azosensors.com/article.aspx?ArticleID=1447.

[66] K. Edward, "Air pollution (chapter 18)," in *Environmental Pollution and Control (Fourth Edition)*, fourth edition ed., J. J. Peirce, R. F. Weiner, and P. A. Vesilind, Eds. Woburn: Butterworth-Heinemann, 1998, pp. 245–269.

[67] U. Ackermann-Liebrich, "Respiratory and cardiovascular effects of NO2 in epidemiological studies," in *Encyclopedia of Environmental Health*, J. Nriagu, Ed. Elsevier, 2011.

[68] Sensortech, https://www.sgxsensortech.com/content/uploads/2014/08/0278_Datasheet-MiCS-4514.pdf.

[69] SPEC Sensor, https://aqicn.org/air/view/sensor/spec/o3.spec-3sp-o3-20.pdf.

[70] CO2Meter, https://www.co2meter.com/blogs/news/gas-sensor-calibration.

[71] C. Houxin et al., "A new calibration system for low-cost sensor network in air pollution monitoring," *Atmospheric Pollution Research*, vol. 12, no. 5, p. 101049, 2021.

[72] S. Devarakonda, P. Sevusu, H. Liu, R. Liu, L. Iftode, and B. Nath, "Real-time air quality monitoring through mobile sensing in metropolitan areas," *ACM SIGKDD'13*, 2013.

[73] L. Spinelle, M. Gerboles, M. G. Villani, M. Aleixandre, and F. Bonavitacola, "Calibration of a cluster of low-cost sensors for the measurement of air pollution in ambient air," in *SENSORS, 2014 IEEE*, 2014.

[74] P. Han et al., "Calibrations of low-cost air pollution monitoring sensors for CO, NO2, O3, and SO2," *MDPI,Sensors*, 2021.

[75] S. Munir, M. Mayfield, D. Coca, A. Jubb, and O. Osammor, "Analysing the performance of low-cost air quality sensors, their drivers, relative benefits and calibration in cities—a case study in sheffield," *Springer, Environmental Monitoring and Assessment*, 2019.

[76] C. Malings et al., "Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring," *Atmospheric Measurement Techniques*, 2019.

[77] I. Patwardhan, S. Sara, and S. Chaudhari, "Comparative evaluation of new low-cost particulate matter sensors," in *2021 8th International Conference on Future Internet of Things and Cloud (FiCloud)*, 2021, pp. 192–197.

[78] R. Bose, A. Parmar, H. Narla, and S. Chaudhari, "Comparative evaluation of low-cost co2 sensors for indoor air pollution monitoring," in *2022 IEEE 8th World Forum on Internet of Things (WF-IoT)*, 2022, pp. 1–6.

[79] C. Nuria et al., "Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?" *Environment International*, 2017.

[80] *Aeroqual Series 500 – Portable Air Quality Monitor*, accessed 8 Feb, 2022, https://www.aeroqual.com/products/s-series-portable-air-monitors/series-500-portable-air-pollution-monitor.

[81] S.-C. Lee and B. Wang, "Characteristics of emissions of air pollutants from burning of incense in a large environmental chamber," *Atmospheric Environment*, 2004.

[82] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, 2021.

[83] CPCB, *National Air Quality Index*, accessed 27 sep, 2022, "https://app.cpcbccr.com/ccr_docs/FINAL-REPORT_AQI_.pdf".

[84] "System of air quality and weather forecasting and research, ministry of earth science, Govt. of India. Indian Institute of Tropical Meteorology, Pune," accessed 12 Feb, 2022, http://safar.tropmet.res.in/AQI-47-12-Details/.