Efficient MLTL Calibration Model for Monitoring the Real-Time Pollutant Emission from Brick Kiln Industry

Sahaya Sakila V.* and Manohar S.

Department of Computer Science and Engineering, College of Engineering and Technology, SRM Institute of Science and Technology, Vadapalani Campus, Tamil Nadu, India
Email: sv5969@srmist.edu.in (S.S.V.); manohars@srmist.edu.in (M.S.)
*Corresponding author

***Abstract*—Coal-ablaze Brick Kiln industries are the major contributors of Particulate Matter ($PM_{2.5}$, PM_{10}) emissions that endanger the environment and pose a variety of health risks to all the living beings. Current static ambient pollutant monitoring stations are sparsely located due to their expensive deployment. Recent advancements in Internet of Things (IoT) technology tends to have portable sensors which could be easily deployed at any location to monitor the quality of air. Calibration for these portable sensors requires training data from static reference monitoring stations. In this study, Brick Kiln industry, which are usually remotely located from the reference stations, is chosen to monitor its emission through the IoT devices, and the calibration for the portable sensors are performed using data from a reference sensor. Calibration of the sensor reading is performed using proposed Meta Learning based Transfer Learning (MLTL) and its performance is evaluated utilizing evaluation metrics of various Machine Learning (ML) and Deep Learning (DL) based regression models. The proposed model shows the most significant scores 0.992236, 0.0002, 0.0048 for the evaluation metrics, R-squared, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE), respectively, as compared to other ML models while calibrating the Particulate Matter (PM) pollutant's emission rate obtained from the industry.**

***Keywords*—brick kiln industry, meta learning-based transfer learning, machine learning, deep learning**

I. Introduction

High levels of air pollution are thought to pose the greatest threat to global environmental health [1]. The two most significant air pollutants worldwide are the Particulate Matter Pollutants $PM_{2.5}$ and PM_{10} [2]. They have a much greater hovering tenacity, exhibit clearly optimized consequences on airborne gases and toxic chemicals, and are primarily responsible over the formation of strontium, resulting in decreased sight at the surface and leads to Low Birth Weight amid kids. They are also capable of permeating the bronchi and alveoli, lung tissues, and blood by means of the nostrils, which can trigger relentless pulmonary diseases, asthma, and cardiovascular glitches [3–6].

Industries are the important sources of Particulate Matter pollutants. Small-scale organizations, like the brick industry, contribute significantly to deteriorating the environment by disseminating substantial quantities of gas and particulate pollution during the season of brick production [2, 7, 8]. During the brick-making period of time, residing 2 km offshore from a brick industry leads to a spike in daily mean $PM_{2.5}$ concentration, that is about five times greater than the World Health Organization (WHO) recommendation for 24-hour exposure (15 $\mu g/m^3$) [9]. The National Air Quality Index (AQI) of the Environmental Protection Agency (EPA) is one of the most widely used methods for evaluating air quality [10].

Typically, the traditional static air pollution monitoring stations which have a high degree of data accuracy and can measure a wide range of contaminants are used to assess the quality of the air. These stations are built with elegant, and time-tested equipments which employs sophisticated measuring techniques, includes auxiliary devices like temperature regulators, relative humidity controllers, air filters, and calibrators to boost the accuracy of the data. As an outcome, the price of a single station frequently exceeds $100,000 which is very expensive to operate and hence sparsely deployed. It also requires frequent service from qualified engineers, provides an inaccurate representation of the city's air quality because they ignore regional differences as most of the metropolitan regions possess a single monitoring station [11–13], and hence it underlines the need for affordable installation which can still measure accurate pollutant rates and be easy to deploy at several different places [14].

The latest developments in wireless networking technologies like the Internet of Things (IoT) and low-cost sensor equipment's provides us the chance to employ arrayed sensor networks to assess the air pollution in real time at a wide range of destinations [10]. They provide a cost-effective alternative due to their increased mobility, smaller size, lesser demands on upkeep, and offers scalability by aiding in the establishment of omnipresent monitoring networks, which will address the problem of

Manuscript received August 8, 2023; revised August 30, 2023; accepted October 12, 2023; published April 28, 2024.

geographical sparseness that the current network of static air monitoring stations faces. Among all the available less expensive technologies, they use, nephelometry, Optical Particle Counters (OPC), Non-Dispersive Infrared (NDIR), Metal Oxide Semiconductor (MOS), Electrochemical Cell (EC), and Light-Scattering Particle Sensors (LSPS) which is the most predominantly deployed technique in existing in-expensive sensors [1, 13, 15, 16].

These sensors are, however, known to be somewhat inaccurate and less dependable than static monitoring stations due to characteristics like weak repeatability, susceptibility to cross-sensitivities between various pollutants, frequent recalibration prerequisite, fluctuation in metrics with altering ambient conditions like temperature and humidity, limited life spans, potential for interference from traffic and weather oscillations, and lack of maturity [11, 13, 15]. Also, Concas *et al.* [13] investigated the impacts of temperature and humidity on inexpensive Particulate Matter (PM) sensors and discovered that relative humidity had an impact on sensor accuracy, but temperature had no discernible effect but it does affect the concentration of particles, which implies, calibration is required on the field to obtain a reliable and accurate sensor reading [1]. In this study, it focuses on the calibration of sensor data to provide better reliability and accuracy of the measured pollutant readings while monitoring the air quality near the Brick Kiln industry.

The first constraint or difficulty in this study's attempt to examine the quality of atmospheric air near brick industries is the use of stationary air examining stations, where samples of the air are only available from a small number of places and are only taken once every hour. Additionally, building and maintaining it cost a lot of money and manpower. Second, the SDS011 sensor used in this study is responsive to atmospheric factors such as temperature and humidity, and as a result, the maximum temperature and relative humidity for which the data are valid are 50 °C and 70%, respectively. As a result, calibrating the sensor values is necessary. To do this, the proposed Meta Learning based Transfer Learning (MLTL) is used. The effectiveness of MLTL is assessed using assessment metrics from various Machine Learning (ML) and Deep Learning (DL) based regression models.

The following section of the article are structured as: Section II discusses several academic works and drawbacks of the current strategy. Section III describes the research area identified for monitoring the quality of air. The proposed MLTL model is described in Section IV, along with the calibration procedures used to validate the sensor readings, the distinct resources and techniques needed to complete the sensor calibration are thoroughly explained in Section V. Section VI describes the various calibration techniques for calibrating the sensor data using ML and DL algorithms. Sections VII and VIII describes the results obtained and model evaluation using various metrics and the study's conclusions, difficulties encountered, and potential for future development are discussed in Section IX.

II. Related Work

The traditional air monitoring stations, though, are very accurate and can measure wide range of contaminants, they are present in limited locations as they have huge infrastructure, costly to deploy, and require qualified engineers for frequent servicing. Hence, the presence of low-cost sensors which are easy to deploy at any needed sites, provides mobility, offers scalability, smaller infrastructure and easy to maintain are highly good at measuring PM pollution from brickyards which are situated far away from the presence of static air monitoring stations. However, they are sensitive to external factors in the environment like temperature, humidity, prone to unreliable readings, and unpredictable working conditions [12]. Hence, calibrating these systems is essential for evaluating and affirming their performance [14, 17].

Traditionally, these devices are calibrated by using conventional techniques such as moving average calculations or by affirming the readings with publicly accessible reference equipment. Nevertheless, this could result in an inaccurate index at the site of the measurement as well as an excessive smoothing of the signal [8, 18]. Calibration based on ML outperforms the traditional procedures, but it necessitates the deployment of a reference monitor along with a significant quantity of training data from the sensor. A variety of statistical techniques like ARIMA, Kalman filtering, and trained ML techniques including linear regression, Nearest Neighbors, and Support Vector Regression (SVR) have been investigated by Yadav *et. al.* [1] and Malyan *et. al.* [19].

For field calibration, the constructed sensor nodes were placed along with a precise reference sensor [10]. In comparison to the data from the reference sensor, Multi Linear Regression (MLR) based temperature and humidity rectification produced Mean Absolute Percentage Error (MAPE) of 48.71% and an R^2 of 0.607. With a MAPE of 38.89% and an R^2 of 0.78 in comparison to the data from the reference sensor, Artificial Neural Network (ANN) based calibration has the potential for substantial additional improvement. PM_1 and $PM_{2.5}$ calibration using and Random Forest Regression (RFR) techniques, respectively produced excellent results, however, PM_{10} calibration using RFR technique needed significant improvement.

On-field calibration models for PM pollutants were performed through ML techniques using MLR, SVR, Gradient Boost Regression (GBR) algorithms and the assessment findings showed a substantial boost in the accuracy (sometimes exceeding up to $R^2 > 0.9$) for the PM sensors that are based on light scattering technology [13]. Air quality monitoring for Carbon Monoxide (CO) and Nitrogen dioxide (NO_2) pollutants were calibrated using MLR, RFR and ANN techniques and the evaluation metrics had shown significant improvement as R^2 was greater than 0.9 for CO in all the three techniques whereas it needed much improvement as R^2 dropped to even lesser than 0.7 for NO_2 pollutant while using RFR technique [15, 20].

TABLE I. INVESTIGATIONS ON DIFFERENT IOT RESOURCE CONSTRAINTS

Ref No.	Methodology	Measured Pollutants	Summary of Contribution
Das *et al.* [21]	Low-cost-Innovative-Air-Pollution-Monitoring-Device (LCI-APMD)	$PM_{2.5}$, PM_{10}, Carbon monoxide (CO), Sulphur dioxide (SO_2), Nitrogen dioxide (NO_2), Ozone (O_3)	As IoT devices have limits in power usage, the study has proposed LCI-APMD, where Particulate Matter and electrochemical sensors are installed and accessed versus a precise reference configuration. It was apparent that the implemented approach maintained a tolerable level of detecting errors while being 91% more power efficient than the precise reference configuration and having a much greater coverage area.
Zaidan *et al.* [22]	LCS-MLM-VS (Low-cost-sensors integrated with ML calibration models and virtual sensors)	$PM_{2.5}$, Carbon dioxide (CO_2)	The LCS-MLM-VS approach demonstrates how to calibrate $PM_{2.5}$ and CO_2 pollutants using non-linear ML models and Virtual Sensors, respectively. Since it is not possible to calibrate the LCS of CO2 using ML models, mathematical models known as VS are established. As a result, LCS can be independently implemented in the field with high precision, which encourages scaling-up precise air pollution mapping suitable for smart cities.
Li *et al.* [23]	RCM-HD (Robust Calibration Methodology using Historical Data)	NO_2 (Nitrogen dioxide)	The low-cost sensors are calibrated using historical information by the RCM-HD methodology, and the sensor slip is subsequently corrected by adjusting their sensitiveness and drift based on the concentration dispersion of the pollutant. RCH can improve the precision and uniformity of low-cost air quality sensors with no aid of real-time and close by data sources, according to analysis in this study, and it performs similarly to field calibration techniques currently in use that use spatially adjacent real-time references.
Ghosh *et al.* [24]	HCSS-MSN (Autocalibration methodology of low-cost Miniature Sensing Nodes, with the help of High-Cost Sensing Stations)	PM (Particulate Matter)	The HCSS-MSN technique uses an appropriate ML regression model to automatically calibrate the data from the low-cost sensors with the information gathered from the HCSS. This study suggests a design for a low-cost, low-power PM sensor to achieve auto-calibration and avoid the need to take MSN offline to calibrate or recalibrate. This sensor was found to be 91% more affordable and 57% more energy-efficient than the commercial, high-cost PM sensor, while still keeping its detection error inside a predetermined limit.
Ali *et al.* [25]	LC-ECS-IR (Low-Cost Electrochemical Sensor and Infrared sensor)	CO (Carbon monoxide), NO_2 (Nitrogen dioxide) and PM (Particulate Matter)	In this article, an LC-ECS-IR has been developed that measures CO and NO_2 concentrations using inexpensive electrochemical sensors and PM levels with an infrared sensor. For field calibration, the created sensor nodes were placed close to an exact standard CO sensor. Data from the low-cost sensor that had been offset and gain calibrated had high agreement with data gathered from the standard sensor. In comparison to the data from the standard sensor, the Multiple Linear Regression-based calibration model had a Mean Absolute Percentage Error (MAPE) of 48% and an R^2 of 0.6, but the Artificial Neural Network calibration model had an opportunity for improvement with a MAPE of 38% and an R^2 of 0.8.
Apostolopoulos *et al.* [26]	AQ-ENSENSIA (An air quality monitoring device developed in the Institute of Chemical Engineering Sciences, Greece)	NO_2 (Nitrogen dioxide) and O_3 (Ozone)	The reference apparatus was employed as the assessment standard in this study as it looked into several methods for the field calibration of NO_2 and O_3 measured by the ENSENSIA air quality monitoring system. For two years, the sensors were placed in the same places. Several Machine Learning (ML) and Deep Learning (DL) algorithms were trained using data from the first year (2021) of seven ENSENSIA sensors (NO_2, NO, O_3, CO, $PM_{2.5}$, temperature, and relative humidity), and the resulting calibration algorithms were evaluated using details from the following year (2022). The O_3 and NO_2 calibration performed best with the Random Forest algorithm, with mean errors of 4.3 ppb and R^2 of 0.69 for O_3 and 3 ppb and R^2 of 0.86 for NO_2 respectively.

III. STUDY AREA

In India, particularly in southern districts, Brick kilns that burn coal have proliferated quickly and are one of the leading causes of air pollution. Despite the widespread documentation of the detrimental impacts of air pollution, there is scant concrete information on the repercussions of this crucial industry. It was observed that there are cluster of brick industries around Aralvaimozhi, a small town in Kanyakumari, Tamil Nadu as captured in the Fig. 1. Field study was conducted at a brick kiln in Aralvaimozhi to measure the $PM_{2.5}$ and PM_{10} pollutants emitted through them in the process of brick-making.

Various environmental factors such as temperature, humidity, wind speed and direction, precipitation, atmospheric pressure, dust and pollen, and geographical locations tend to have an impact on the measured PM pollutants. In this study, calibration of measured PM readings is performed against its sensitivity towards temperature and humidity.

IV. PROPOSED SYSTEM

The process of meta-learning entails teaching a model how to learn. When referring to sensor calibration, it refers to teaching a meta-model to swiftly adjust to various calibration jobs. Various steps involved in the Meta-Learning processes are, initially, the data input layer receives input data from SDS011 and DHT11 in the form of csv file, which includes time-stamp, sensed PM pollutant readings, AQI of the corresponding PM pollutants, temperature and humidity data. When referring to meta-learning-based calibration, the term “ground truth” describes the information from the reference sensor (DHT11) that is utilised as a training signal to instruct the model how to carry out calibration. This information acts as a stand-in or replacement for the actual ground truth

values, which could be acquired through more sophisticated and pricey reference equipment. A fixed-size feature vector is created by a neural network after it learns to gather pertinent data from the SDS011 and DHT11 sensors. The meta-learner gains the ability to correspond the proper calibration parameters with the input data (SDS011 and DHT11 measurements). To ensure that the meta-learner generalizes successfully to various environmental variables, this procedure is performed for a variety of scenarios. The meta-learner does not formally "identify" faulty sensor readings. Instead, it discovers from instances and patterns in the training data that there is typically a constant difference between SDS011 and DHT11 values when temperature and humidity surpass particular thresholds. In order to increase the accuracy of SDS011 measurements in real-time, even when traditional ground truth values for faulty situations are not available, it learns to forecast calibration modifications based on these observed associations.

Transfer learning is a strategy that involves transferring or adapting knowledge obtained from a meta-learner (Origin location) to enhance performance on a target location. The next stage is to transfer this knowledge to the transfer learning algorithm when the origin location has completed training and acquired knowledge about how to calibrate the SDS011 sensor. Transfer learning enables the meta-learner's calibrated features and insights to be used in future calibration tasks utilizing the SDS011 sensor. Various steps involved in the Transfer-Learning processes are, the meta learner (Origin Location) has already frozen the feature selection process, and the target location (Target Location) receives the knowledge directly. The transfer learning approach refines the model using the new dataset based on the knowledge learned from the origin location. For any new input received from SDS011 and DHT11 sensors, it performs calibration by modifying the learned features and calibration information from the meta-learner. Thus, the sensor's accuracy and adaptability to different environmental circumstances are improved by the Meta Learning based Transfer Learning (MLTL) approach, even in situations where reference data may be scarce or nonexistent.

Any learning algorithm are enhanced through the process of "Meta Learning," commonly referred to as "acquisition of knowledge to acquire knowledge," so that it may quickly pick up novel abilities or cope with unfamiliar settings with only a modest amount of data used as training. It seeks to hone a model across a variety of challenges in order to develop a versatile approach that is capable of being versatile across tasks, perhaps even unidentified ones. Every task performed has a related S and $\int_S$, which represents the data collected and a loss function, respectively. Consequently, obtaining the most ideal variables for the model becomes the mantra for ML, which are holistically assessed as in Eq. (1).

$$ð^* = \arg\min_{ð} E_{S\sim p(S)} [\int_S (ð)] \quad (1)$$

where, p(S) is the likelihood spread among the datasets. Any sort of model that evolves via gradient descent is interoperable with MLTL, which is a very universal meta learning approach based on optimization. The algorithm's key principle is to acquire the model's initial variables so that it may be efficiently modified for optimum efficiency on a fresh assignment. It is a dual-step approach to optimization that entails training a base model after acquiring a meta-learner. The meta-learner looks for the best initial variables for each base model dependent on its task, to ensure that the base model could be coached rapidly with minimal data.

A. Calibration of Sensor through MLTL

The proposed MLTL approach of meta-learning that leverages the data from multiple origin locations, to coach the base model that is positioned at the target site by utilizing the scant training data available there. Meta learner is universal across all the sites, whereas the base model is subjected to appropriate sites. In this study, a function $\boldsymbol{\theta}_{Si}$ with a parameter $\boldsymbol{\theta}_{Si}$ are used to represent the base model, that is executed as a completely integrated neural network. It uses the assistance from meta learner which is similarly integrated completely with the neural network, represented by fð with parameter ð, to efficiently and swiftly coach the base model at any site s. The fundamental structure of fð and $f\theta_s$ is identical. Fig. 1 represents the Details of Proposed Calibration Model. To begin with, the meta learner is coached using the origin data, where the data are split into Prefer Array and Reserve Array from the origin location S^0. The Prefer Array are helpful to leverage the different variables present in the base model, whereas the Reserve Array helps in assessing the base model's efficiency. Followed by this, the base model is coached using the restricted coaching data available through meta learner and target site. The $f\theta_s$ of base model is gets launched through the meta learner fð, as explained in Fig. 1

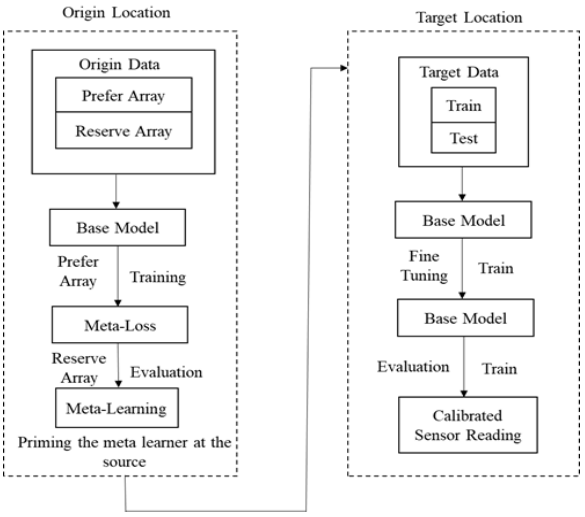


Fig. 1. Proposed Meta Learning-based Transfer Learning (MLTL) model for calibrating the sensor readings.

B. Priming the Model at Origin Location

The fð of meta learner and $\boldsymbol{\theta}S_i \leftarrow ð$ of base model gets launched erratically. The base model is coached through gradient descent represented as, where $\lambda \in R$ is the rate of

learning. MAE evaluation metric is used by the loss function.

$$\boldsymbol{\theta}_{Si} \leftarrow \boldsymbol{\theta}_{Si} \leftarrow \lambda \, \Delta_{\boldsymbol{\theta}Si} \int_{T_{Si}^{P}} (\boldsymbol{\theta}_{Si}) \quad (2)$$

The base model is assessed through the Reserve Array by determining loss $\int_{T_{Si}^{R}} (\boldsymbol{\theta}_{Si})$. At all the origin locations, the complete process is put on loop as shown in Fig. 1. By limiting the meta loss, the meta learner is coached, which is the entirety of all the distinct losses $\int_{T_{Si}^{P}} (\boldsymbol{\theta}_{Si})$, performed with the reserve array, followed by the process of optimizing.

$$ð \leftarrow ð - \mu \sum_{i} \Delta_{ð} \int_{T_{Si}^{R}} (\boldsymbol{\theta}_{Si}) \quad (3)$$

where learning rate is defined by $\mu \in R$. At the end, the base models from various origin sites are dumped out and the meta learner alone is preserved.

C. Priming the Model at the Target Location

The model is launched again with ð and in Eq. (2). and Eq. (3). are utilized to coach θS_j, to better forecast at the target site $S_j \in S^t$, with little coaching data. The train set performs as the coaching data, but the efficiency of calibration is evaluated by the test set. The pseudocode of the complete coaching algorithm is pictured in the Algorithm 1.

Algorithm 1. MLTL Based PM Sensor Reading Calibration with reference sensor device

Input: S^o → set of origin location, S^t → set of target location, λ, μ: hyperparameters
1: erratically initiate ð
2: while not performed do
3: Represent a set of location from S^o
4: for sensor deployment location S_i do
5: Initiate the base model $\boldsymbol{\theta}S_i \leftarrow ð$
6: Represent prefer Array $T_{Si}^{P} = \{X_s, Y_s\}$ from S_i
7: Assess $\Delta_{\boldsymbol{\theta}Si} \int_{T_{Si}^{P}} (\boldsymbol{\theta}_{Si})$
8: Calculate parameter with gradient descent optimizer: $\boldsymbol{\theta}_{Si} \leftarrow \boldsymbol{\theta}_{Si} \leftarrow \lambda \, \Delta_{\boldsymbol{\theta}Si} \int_{T_{Si}^{P}} (\boldsymbol{\theta}_{Si})$
9: Sample Reserve set: $T_{Si}^{R} = \{X_s, Y_s\}$
10: Evaluate $\int_{T_{Si}^{R}} (\boldsymbol{\theta}_{Si})$
11: end for
12: reform $ð \leftarrow ð - \mu \sum_{i} \Delta_{ð} \int_{T_{Si}^{R}} (\boldsymbol{\theta}_{Si})$
13: end while
14: for target location S_j £ S^t do
15: Initiate the model with $\boldsymbol{\theta}_{Sj} \leftarrow ð$
16: Calculate the prime parameters θ_{Sj}^{*} with gradient descent on the prefer /priming set: $\boldsymbol{\theta}_{Sj} \leftarrow \boldsymbol{\theta}_{Sj} \leftarrow \lambda \, \Delta_{\boldsymbol{\theta}Sj} \int_{T_{Sj}^{P}} (\boldsymbol{\theta}_{Sj})$
17: Assess on Reserve/test Array
18: end for
output: Calibrated Sensor Reading

V. Materials and Methods

A. Experimental Design

At the chosen Brick kiln, the necessary gear, including a SDS011 sensor, DHT11 sensor, Raspberry Pi Model 3, cables as well as Wi-Fi or mobile data access is setup to measure PM pollutants in real-time, released by the brick kilns. The SDS011 sensor, which is effective and reasonably priced, is used to measure the PM concentrations emitted near the brick kiln. Temperature and humidity are two atmospheric variables that are measured using a DHT11 sensor, which is incredibly energy-efficient, relatively simple to set up and use, allows for blending into different devices and systems, streamlines data acquisition and processing, offers real-time data on temperature and humidity, enables applications to constantly monitor and react to variations in environmental conditions, operates at low voltage levels, and uses little power while providing accurate temperature and humidity measurements. After receiving a power source, the SDS011 and DHT11 sensors are connected to a Raspberry Pi Model 3 via UART and GPIO pins, respectively to gather data on the concentration of PM pollutants, temperature, and humidity. A Wi-Fi or mobile data source is also enabled on the Raspberry Pi model. As shown in the Fig. 2, the DHT11 sensor's ground connection is attached to GPIO pin 6 via a brown wire, its data pin is connected to GPIO pin 4 via a white cable, and its power connector (5V power supply) is connected to the Raspberry Pi's physical pin 2 for power via a black cable. The real-time sensor data were collected between May 2023 and June 2023 from the Brick Kiln Industry, where the SDS011 and DHT11 sensors recorded values for every 60 s. The data is preserved as a CSV file with information including the date and time stamp, the $PM_{2.5}$ and PM_{10} concentrations, the AQI values for the two concentrations, the humidity, and the temperature. The unprocessed data from the SDS011 sensor is transformed into AQI data, as indicated in section V-C. The air quality will be assessed using the resulting AQI data, where the calibration of sensor data is carried out utilizing ML and DL algorithms to make the data more reliable and accurate.

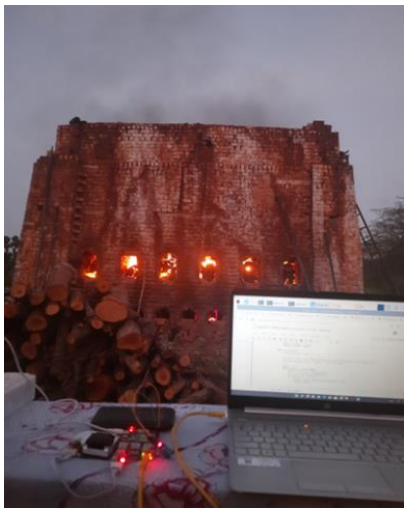


Fig. 2. Picture of the deployment set up, to examine the quality of atmospheric air in brick kiln industry.

B. Data Acquisition

The Nova SDS011 PM Sensor is employed for identifying the PM pollutants since it offers the best

price/accuracy ratio. With a detection range of 0.3 to 10 μm, it takes advantage of the light scattering theory. DHT11 is considered as the reference sensor in this study, which is effective in measuring atmospheric factors like temperature and humidity.

Data collection involves gathering a dataset made up of pairs of sensor measurements and related values of reference. By using an SDS011 sensor, real-time PM concentration emitted from the Brick kiln are gathered and stored as a csv file. As this sensor is sensitive to atmospheric factors such as temperature, when it is greater than 50 ºC, and humidity, when it is greater than 70%, the SDS011 sensor readings obtained above these threshold values, in this study, are trained to be calibrated using ML and DL algorithms to avoid falsification of data and to provide accurate and precise level of PM concentrations emitted from the Brick Kiln industries. In Fig. 3, the entire work flow for data collection and calibration of sensor readings is described.

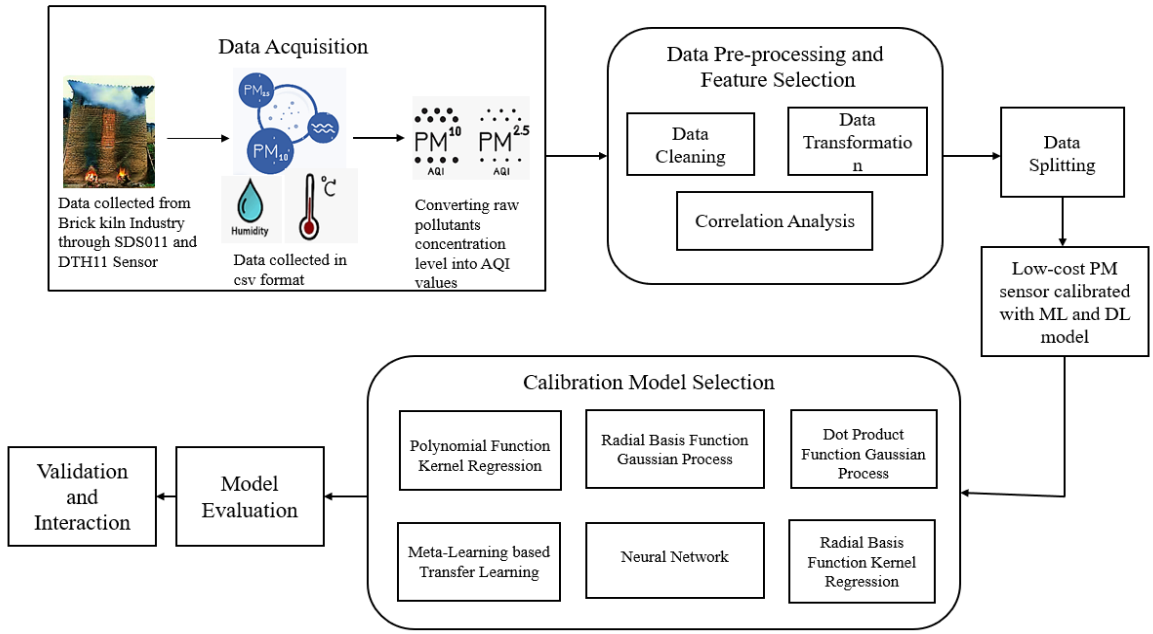


Fig. 3. The proposed workflow for monitoring particulate matter pollutants emitted in brick kiln industry.

C. *Decrypting the Raw Pollutant Concentration Data*

A tool for gauging air quality is the Air Quality Index (AQI). As indicated in Table II, it is an evaluation scale which is devoid of units, which serves in the purpose of rating the quality of air as Good, Satisfactory, Moderate, Poor, Very Poor and Severe for breathing. It is a value derived from the Eq. (4).

$$AQI(PM_{2.5,10}) = \left[\left\{\frac{(\emptyset_{max}-\emptyset_{min})}{(\theta_{max}-\theta_{min})}\right\} \times (\mu - \theta_{min})\right] + \emptyset_{min} \quad (4)$$

where,

$AQI(PM_{2.5,10})$ = Particulate Matter's Air Quality Index

θ_{max} = Highest threshold value in accordance to pollutant's raw data

θ_{min} = Lowest threshold value in accordance to pollutant's raw data

$\emptyset_{max}$ = Highest AQI value in accordance to θ_{max}

$\emptyset_{min}$ = Lowest AQI value in accordance to θ_{min}

μ = Pollutant's raw concentration rate

Few instances like, if the sensor measures the concentration of $PM_{2.5}$ as 58 μg/m^3, its AQI will be calculated from Eq. (4) as, $AQI_{PM2.5} = \left[\left\{\frac{100-51}{60-31}\right\} \times (58 - 31)\right] + 51$, which is approximately estimated to be 97, and falls within the "Satisfactory" category as mentioned in Table II, which implies that the level of air quality is deemed adequate, with only a very small chance that those with exceptionally high sensitivity to air pollution would have cause for worry and minor health consequences may be experienced by sensitive populations.

If the sensor measures the concentration of $PM_{2.5}$ as 189 μg/m^3, its AQI will be calculated from Eq. (4) as, $AQI_{PM2.5} = \left[\left\{\frac{400-301}{250-121}\right\} \times (189 - 121)\right] + 301$, which is approximately estimated to be 353, and falls within the "Very Poor" category as mentioned in Table II, which impacts the people who are members of sensitive groups to have a higher risk of suffering worse health effects and has a greater chance that adverse health effects may affect the entire population.

If the sensor measures the concentration of PM_{10} as 42 μg/m^3, its AQI will be calculated from Eq. (4) as, $AQI_{PM10} = \left[\left\{\frac{50-0}{50-0}\right\} \times (42 - 0)\right] + 0$, which is approximately estimated to be 42, and falls within the "Good" category as mentioned in Table II, which implies that the air quality is in a favorable range and presents either no risk or a negligible risk to health and this condition is optimal for engaging in outdoor activities and exercises.

If the sensor measures the concentration of PM_{10} as 285 µg/m^3, its AQI will be calculated from Eq. (4) as, $AQI_{PM10} = \left[\left\{\frac{300-201}{350-251}\right\}\times(285-251)\right]+201$, which is approximately estimated to be 235, and falls within the "Poor" category as mentioned in Table II, which implies that there is a possibility that the general population might have health impacts as a result of the deteriorating air quality and the members of vulnerable groups can suffer worse health effects.

Likewise, if the calculated AQI falls under "Moderate" category, it implies that the air quality is somewhat polluted and could impact individuals in sensitive groups, those with respiratory or heart conditions might encounter health effects, while the general public is relatively less susceptible to these effects. And, if the calculated AQI falls under the "Severe" category, it implies that there is an increased likelihood that the entire population will be impacted, and more severe health effects could be experienced by everyone.

TABLE II. AQI BASED $PM_{2.5}$ AND PM_{10} CONCENTRATION BREAKPOINT GIVEN BY CENTRAL POLLUTION CONTROL BOARD (CPCB)

AQI Range	Observation	PM_{10} (24 h)	$PM_{2.5}$ (24 h)	Colour Code
0–50	Good	0–50	0–30	
51–100	Satisfactory	51–100	31–60	
101–200	Moderate	101–250	61–90	
201–300	Poor	251–350	91–120	
301–400	Very Poor	351–430	121–250	
401–500	Severe	430+	250+	

D. Data Preprocessing and Feature Selection

The dataset is first cleaned by dealing with outliers, missing numbers, and any other data quality concerns. Reference values of temperature and humidity are regarded as independent variables since it is expected that they will affect the readings from the SDS011 sensor. Only essential features that are required for regression analysis, such as atmospheric variables temperature and humidity, PM AQI values, time stamps, dates, etc., are taken into account during the calibration process. Any other irrelevant or non-informative features from the dataset are excluded. Then, in order to eliminate bias in the regression model caused by variations in magnitude, the SDS011 sensor data, temperature, and humidity variables are standardized to a single scale. The whole collection of data that is required for calibration is structured correctly, and it has been verified that it is compatible for regression analysis.

E. Regression Training Algorithms

To rectify or adjust unprocessed sensor readings, sensor calibration involves forecasting numerical values. Regression models are a suitable choice because they produce continuous numerical data and are created expressly for this use. They are designed to provide accurate estimates across the entire range of sensor values, handle any varying scales and ranges without the need for extensive preprocessing, and provide a direct measure of error or residual, all of which are critical in calibration tasks where precision matters. They are also less sensitive to class imbalances and data distribution, ensuring stable performance even when dealing with varying data patterns. A comparison study which shows the better methodology to calibrate sensor readings using ML methods is discussed in Table III.

TABLE III. MERITS AND DEMERITS OF VARIOUS METHODOLOGIES USED FOR CALIBRATION

S.No.	Methodologies	Description
1	Regression Algorithms	Outputs continuous numerical data, allowing for accurate mapping of sensor readings. Regression models, which are crucial for calibration with numerous variables, can capture complex correlations between sensor readings and physical values. Gives quantifiable error measurements (such R^2, RMSE, and MAE) for measuring calibration accuracy. More resistant to anomalies and erratic data. Allows for continuous calibration and updates and modifications to the model as new data is gathered.
2	Classifier Algorithms	Outputs Discrete class labels for classification tasks may not readily correspond to the numerical range of sensor data, restricting interpolation capabilities. Calibration activities don't immediately relate to classification error measures like accuracy and F1-score. Less capable of capturing complex interactions, especially when nonlinear modelling is needed for sensor calibration.
3	CART (Decision Trees)	Outputs Discrete classes or continuous values, according to the job, but because decision trees divide data into discrete regions, they may not efficiently capture continuous relationships. The accuracy of the calibration may not line up with error metrics since they may be related to classification accuracy or Gini impurity. Complex relationships might not be modelled well, especially when a linear divide is insufficient. Decision tree topologies offer interpretability, but deep trees can lead to complexity.

After performing data cleaning, critical feature selection, and normalization, the entire data is divided into 2 data sets, for coaching and evaluation. In this study, 70% of the data is allocated for coaching and 30% of it is allocated for evaluating. The regression model will be created using the training data, and its effectiveness will be assessed using the test data set. Various regression algorithms like Radial Basis Function Gaussian Process Regression (RBF-GPR), Dot Product Gaussian Process Regression (DP-GPR), Radial Basis Function Kernel Regression (RBF-KR), Polynomial Function Kernel Regression (PF-KR). Neural Network (NN), and Meta Learning based Transfer Learning (MLTL) are used to calibrate the SDS011 sensor reading obtained from Brick Kiln industry.

F. Performance Indexes

Metrics like Mean Square Error (MSE), Root Mean Square Error (RMSE), and coefficient of determination (R^2) are used to evaluate the effectiveness of the regression model. The model's ability to depict the link between the sensor readings and the independent variables is evaluated. Then, depending on the assessment measures, the most effective model is chosen, and it is then utilized to calibrate the fresh set of sensor values.

VI. Calibration Techniques

A. Gaussian Process Regression (GPR)

Two kernels are utilized with Gaussian Process regression, namely, Radial Basis Function (RBF) and Dot product (linear) Kernel. GPR forecasts a range of probabilities across potential functions rather than an individual point projection, allowing for the measurement of prediction ambiguity.

1) Radial Basis Function Gaussian Process Regression (RBF-GPR)

RBF, also referred as squared exponential kernel, assesses the degree of resemblance across data points in the feature space depending on their Euclidean distance. It functions as based on the Eq. (5).

$$RBF_k(D_x, D_y) = \exp\left(-\frac{\|D_x - D_y\|^2}{2\omega^2}\right) \quad (5)$$

where, D_x *and* D_y are the datapoints, Euclidean distance between the data points is defined as $\| D_x - D_y \|$, and ω represents the bandwidth parameter. Smooth patterns in the data can be captured effectively by the RBF kernel.

2) Dot Product Gaussian Process Regression (DP-GPR)

Dot Product, also referred as Linear kernel, assesses the resemblance of two data points by looking at the inner product of each. As it is represented in Eq. (6):

$$Linear_k(D_x, D_y) = D_x^T . D_y \quad (6)$$

where, D_x^T and D_y are the respective transposed and original vectors of the two data points. It is based on the data points' absolute placements inside the feature space and helpful when it is assumed that the characteristics and the target variable are connected linearly to form the underlying function.

B. Kernel Regression (KR)

Two kernels are utilized with Kernel Regression, namely, Radial Basis Function (RBF) and Polynomial Kernel. RBF kernel, as explained in Eq. (5) is commonly used in both KR and GPR.

It calculates the degree of resemblance between data points relying on the polynomial of their inner product and can be represented as in Eq. (7).

$$Polynomial_k(D_x, D_y) = (\beta . D_x^T . D_y + \gamma)^d \quad (7)$$

where, D_x^T *and* D_y are the respective transposed and original vectors of the two data points, β is used as a parameter for scaling, γ represents a bias or offset term and d represents the polynomial's degree. The flexibility of the polynomial kernel, which could capture non-linear correlations between features, relies on the polynomial's degree and they are very effective in handling non-linearly separable data.

C. Neural Network (NN)

TensorFlow libraries being the architect of Neural Network model are used in the process of tuning hyperparameters like number of epochs, neurons present in each layer, number of layers that are hidden, and sample size. Relation between each of these parameters could be presented as in Eq. (8).

$$H_n = \frac{\Delta}{\beta \times (\theta_i + \theta_o)} \quad (8)$$

where, H_n represents the number of layers that are hidden, Δ represents the dataset's size, θ_i represents the number of neurons that are given as input, θ_o represents the number of neurons received as output, and β represents the randomly chosen scaling factor between 2 to 10. Various test results revealed that the model produced superior outcomes, when, there were 5 number of hidden layers, 64 neurons for each layer, 750 epochs, and for a sample size of 10. ReLu and adam are used as an activation function to train the model effectively and as an optimizer, respectively.

D. Meta-Learning Based Transfer Learning (MLTL)

A PM sensor along with a reference sensor is installed at the research area. For the sensors installed at a certain location S, $T_s = (X_s, Y_s)$ Where X_s represents the data gathered from the PM sensor's period of time and Y_s stands for the period of time data acquired from the collectively placed reference sensor, respectively. Because Xs and Ys are the same length, they are identified by the symbol |Ts|. In the proposed model, the Origin location is referenced as S^s and the Target location as S^t; "θ" is referenced as a parameter which encodes the output from the findings, and in order to put F_θ into execution, we employ a neural network as described in Fig. 2.

VII. Model Selection

A model with higher R^2, lower RMSE, and lower MAE is considered more accurate and better at predicting the sensor readings based on its sensitive parameters' temperature and humidity values.

A. R-Squared (R^2) Metric

A mathematical measure called R-squared shows how much of the variation in a dependent variable, in this case, sensor readings, could be predicted based on the independent variables' values for temperature and humidity. It is a score between 0 and 1 that indicates how well the model matches the data. An R^2 value of 1 denotes an impeccable fit, whereas a value of 0 shows that the model does not explain any variation. A greater R^2 value suggests, that the model reflects the data more closely.

$$R^2 = 1 - \frac{\sum_{i=1}^{\theta}(\alpha_i - \hat{\alpha}_i)^2}{\sum_{i=1}^{\theta}(\alpha_i - \bar{\alpha}_i)^2} \quad (9)$$

As in Eq. (9), θ is the dataset's number of data samples, α_i is the i^{th} data point's real sensor target value, $\hat{\alpha}_i$ is the i^{th} data point's predicted sensor target value derived using regression model, $\bar{\alpha}_i$ represents the average of the sensor's actual target value.

B. Root Mean Square Error (RMSE) Metric

The average magnitude of the residuals is gauged using the statistic known as RMSE. It calculates the disparity among the initial sensor readings and the projected values by taking the square root of the mean of the squared discrepancies. The dependent variable's (sensor readings) units are used to express RMSE. Lower RMSE values imply less residuals and a better fit of the model to the data.

$$RMSE(\alpha, \hat{\alpha}) = \sqrt{\frac{1}{\theta}\sum_{i=1}^{\theta}(\alpha_i - \hat{\alpha}_i)^2} \quad (10)$$

C. Mean Absolute Error (MAE) Metric

MAE estimates the average size of the residuals but ignores the squared difference. Similar to RMSE, MAE is measured in the same units as the sensor values that are the dependent variable. A mean absolute measure of errors in forecasting is provided by MAE. Compared to RMSE, it is less susceptible to outliers.

$$MAE(\alpha, \hat{\alpha}) = \frac{1}{|\theta|}\sum_{i=1}^{\theta} |\alpha_i - \hat{\alpha}_i| \quad (11)$$

VIII. Result and Discussion

In this study, with the help of reference sensor device which is capable of recording the atmospheric influencers like temperature and humidity that affect the sensor reading, various Machine Learning and Deep Learning algorithms are used for calibrating the PM sensor data.

Firstly, data obtained from PM sensor and reference device are stored in the form of .csv file which will include sub-items like Date and Time, raw values of $PM_{2.5}$ and PM_{10}, AQI values of $PM_{2.5}$ and PM_{10}, Temperature and Humidity. Calibration is necessary for the sensor readings that exceeds the threshold range of humidity >70% and temperature >50 ºC. Then, the data pre-processing happens as explained in Section IV.C, where the required features with the correlated values are chosen, undergoes one-hot encoding technique for categorical representation based on the threshold ranges of environmental factors and labelled numerically as "0" and "1".

Then, the data is fitted into the various ML and DL based models for the calibration of sensor readings and the results are evaluated based on the evaluation metrics as described in Section VII. Among the various ML and DL models (RBF-GPR, DP-GPR, RBF-KR, PF-KR, NN, and MLTL), Meta-learning-based Transfer Learning model has performed better and given more accurate readings.

As shown in Fig. 4, it achieving low error levels is seen by the training loss gradually decreasing over time. It can be observed that the difference between training loss and validation loss is first noticeably smaller and then progressively expands, illustrating the precision of the model fitting. The performance of different calibration methods is shown in Table IV.

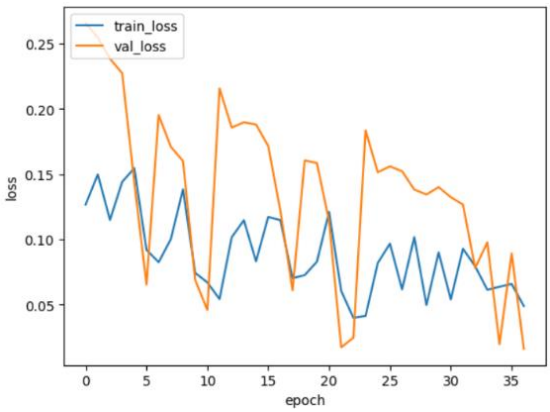
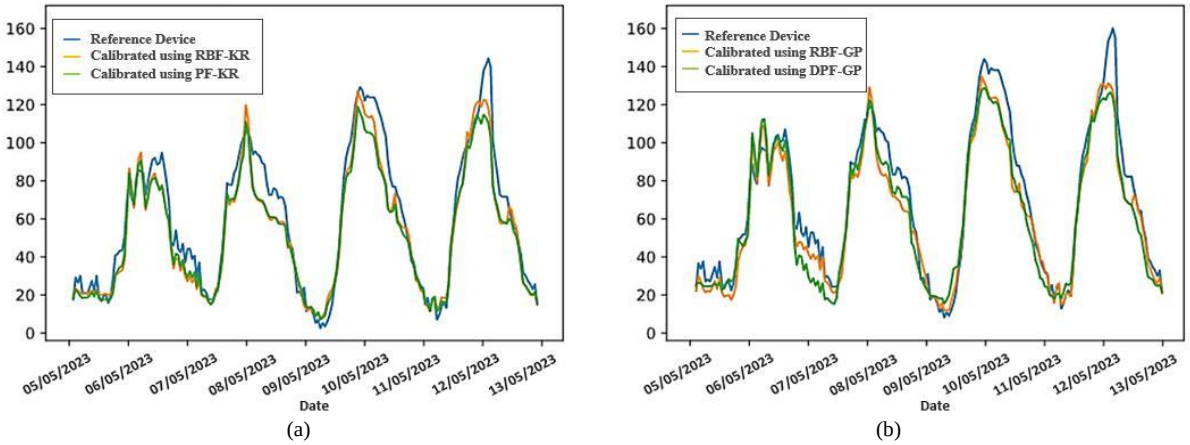


Fig. 4. Estimation of low error rate in MLTL algorithm.

TABLE IV. Calibration Results Obtained by ML and DL Based Algorithms

No	Algorithms	R^2	RMSE	MAE
1	RBF-KR	0.873251	0.0154	0.1196
2	PF-KR	0.914461	0.0053	0.0153
3	RBF-GP	0.940288	0.0102	0.0892
4	DPF-GP	0.935665	0.0368	0.074
5	NN	0.972057	0.0017	0.0045
6	MLTL	0.992236	0.0002	0.0048

Figs. 5–7 illustrates the calibrated sensor readings with the help of reference sensor device through various ML and DL based algorithms.



(a) (b)

Fig. 5. Calibration of sensor data using ML algorithms: (a) RBF-KR and PF-KR; (b) RBF-GP and DPF-GP.

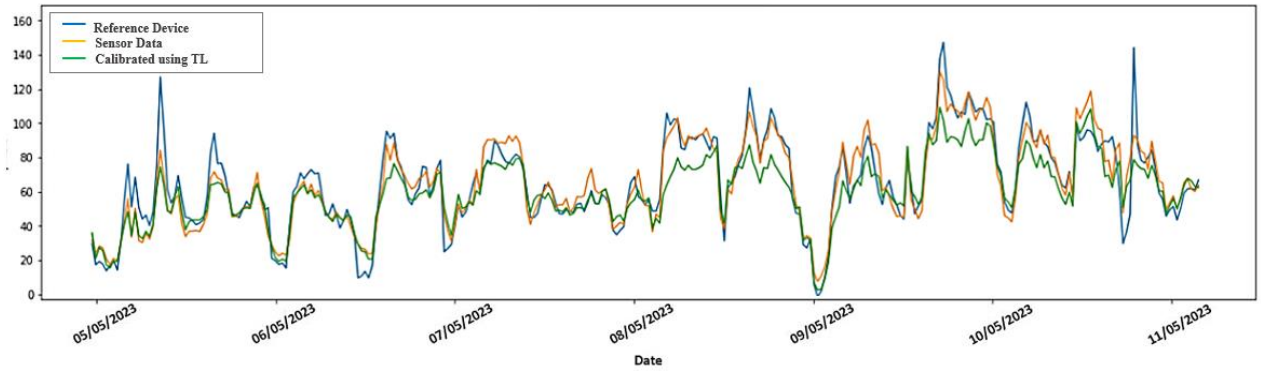


Fig. 6. Calibration of sensor data using MLTL.

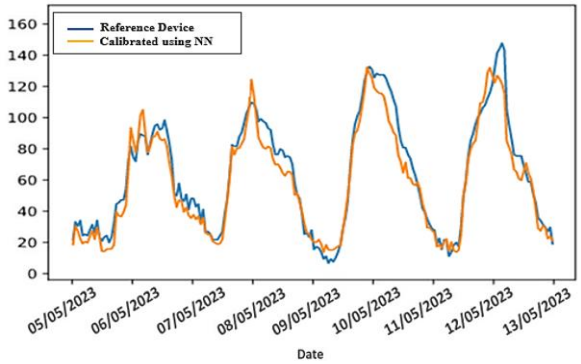


Fig. 7. Calibration of sensor data using NN deep learning based regression algorithms.

IX. Conclusion

In this study, it addresses the effective use of low-cost sensor for monitoring the real-time pollutant's concentration level emitted from the Brick Kiln industry, which is remotely located from the fixed air quality monitoring station. The main contribution of this study is, when there is lack of data from authorized stationary air quality monitoring station to validate the data produced by the air quality monitoring sensors, a high-quality reference device is used to help in the process of calibrating the data obtained from the sensor, by analysing which atmospheric factor has influenced the sensor readings. Once the sensor's sensitive factors are identified, various ML and DL based regression algorithms are used to calibrate the sensor's readings along with the help from reference device to provide accurate and reliable readings. In the process of calibrating the sensor data measured from Brick Kiln industry, among the various regression algorithms used for calibration, the proposed method showed best-in-class performance while evaluating it using all the necessary evaluation metrics followed by regression models. It is evident from this study, that, when there is no data from static monitoring stations which could be used as reference data to validate the sensor's readings, by using this study's reference device's data, the proposed system is capable and very effective in calibrating the sensor data against its sensitivity to environmental factors. In future, proposed method could be extended to calibrate the various other pollutant's concentration level (CO, O_3, SO_2, NO_2, etc.,) that are emitted from industries in real-time, with the help of various IoT protocols. It could be further extended to secure the calibrated data against tampering using Blockchain Technology.

Conflict of Interest

The authors declare no conflicts of interests.

Author Contributions

Sahaya Sakila: Collecting datasets, research concept and methodology, writing original draft preparation. Manohar S: reviewing, editing and validation. All authors contributed to the article and approved the submitted version.

References

[1] K. Yadav *et al*., "Few-shot calibration of low-cost air pollution ($PM_{2.5}$) sensors using meta learning," *IEEE Sensors Letters*, vol. 6, no. 5, pp. 1–4, 2022.

[2] M. S. Minova, S. G. Ilieva, and A. Ivanov, "PM_{10} prediction using CART method depending on the number of observations," in *Proc. 2020 3rd International Conference on Mathematics and Statistics*, 2020, pp. 65–70.

[3] H. Y. Liu, P. Schneider, R. Haugen, and M. Vogt, "Performance assessment of a low-cost $PM_{2.5}$ sensors for a near four-month period in Oslo," *Atmosphere*, vol. 10, no. 2, 41, 2019.

[4] A. R. Fernández *et al*., "Association of prematurity and low birth weight with gestational exposure to $PM_{2.5}$ and PM_{10} particulate matter in Chileans newborns," *International Journal of Environmental Research and Public Health*, vol. 19, no. 10, 6133, 2022.

[5] S. S. Varghese and M. Shanmugavel, "Real-time air quality monitoring in bull Trench kiln-based brick industry by calibrating sensor readings and utilizing the serverless computing," *Expert Systems with Applications*, 121397, 2023.

[6] X. Li *et al*., "Using sensor network for tracing and locating air pollution sources," *IEEE Sensors Journal*, vol. 21, no. 10, pp. 12162–12170, 2021.

[7] V. S. Sakila, S. Manohar, and P. A. Ebenezer, "Ambient particulate matter monitoring system using SDS011 sensor utilizing machine learning approach and ambit of blockchain technology," *Materials Today*, vol. 3, 2023.

[8] V. S. Sakila and A. R. Kavitha, "Analyzing the sources of air pollution and comparing its impact during the phases of COVID-19 pandemic and the scope of IoT in monitoring air quality," in *Proc. International Conference on Communication and Artificial Intelligence: ICCAI 2021*, 2022, pp. 183–197.

[9] N. Brooks *et al*., "Health consequences of small-scale industrial pollution: Evidence from the brick sector in Bangladesh," *World Development*, vol. 170, 106318, 2023.

[10] A. D. Pham *et al*., "Sheaf-theoretic self-filtering network of low-cost sensors for local air quality monitoring: A causal approach," arXiv preprint, arXiv:2212.14353, 2022.

[11] Z. Idrees and L. Zheng, "Low-cost air pollution monitoring systems: A review of protocols and enabling technologies," *Journal of Industrial Information Integration*, vol. 17, 100123, 2020.
[12] M. L. Aix, S. Schmitz, and D. J. Bicout, "Calibration methodology of low-cost sensors for high-quality monitoring of fine particulate matter," *Science of The Total Environment*, 164063, 2023.
[13] 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, 2022.
[14] K. Sridha *et al*., "A modular IOT sensing platform using hybrid learning ability for air quality prediction," *Measurement: Sensors*, vol. 25, 100609, 2023.
[15] H. Chojer *et al*., "Can data reliability of low-cost sensor devices for indoor air particulate matter monitoring be improved—An approach using machine learning," *Atmospheric Environment*, vol. 286, 119251, 2022.
[16] S. D. Vito *et al*., "Adaptive machine learning strategies for network calibration of IoT smart air quality monitoring devices," *Pattern Recognition Letters*, vol. 136, pp. 264–271, 2020.
[17] D. Suriano and M. Penza, "Assessment of the performance of a low-cost air quality monitor in an indoor environment through different calibration models," *Atmosphere*, vol. 13, no. 4, 567, 2020.
[18] E. M. Rueda *et al*., "Size-resolved field performance of low-cost sensors for particulate matter air pollution," *Environmental Science and Technology Letters*, vol. 10, no. 3, pp. 247–253, 2023.
[19] V. Malyan, V. Kumar, and M. Sahu, "Significance of sources and size distribution on calibration of low-cost particle sensors: Evidence from a field sampling campaign," *Journal of Aerosol Science*, vol. 168, 106114, 2023.
[20] P. Souza *et al*., "An analysis of degradation in low-cost particulate matter sensors," *Environmental Science: Atmospheres*, vol. 3, no. 3, pp. 521–536, 2023.
[21] P. Das, S. Ghosh, S. Chatterjee, and S. De, "A low-cost outdoor air pollution monitoring device with power controlled built-in PM sensor," *IEEE Sensors Journal*, vol. 22, no. 13, pp. 13682–13695, 2022.
[22] M. A. Zaidan *et al*., "Intelligent calibration and virtual sensing for integrated low-cost air quality sensors," *IEEE Sensors Journal*, vol. 20, no. 22, pp. 13638–13652, 2020.
[23] G. Li *et al*., "RCH: Robust calibration based on historical data for low-cost air quality sensor deployments," in *Proc. 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and ACM International Symposium on Wearable Computers*, 2020, pp. 650–656.
[24] S. Ghosh, P. Das, S. De, S. Chatterjee, and M. Portmann, "Local reference-free in-field calibration of low-cost air pollution monitoring sensors," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–13, 2022.
[25] S. Ali *et al*., "Low-cost sensor with IoT LoRaWAN connectivity and machine learning-based calibration for air pollution monitoring," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2020.
[26] I. D. Apostolopoulos, G. Fouskas, and S. N. Pandis, "Field calibration of a low-cost air quality monitoring device in an urban background site using machine learning models," *Atmosphere*, vol. 14, no. 2, 368, 2023.

Copyright © 2024 by the authors. This is an open access article distributed under the Creative Commons Attribution License (CC BY-NC-ND 4.0), which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.