Environ Monit Assess (2024) 196:621
https://doi.org/10.1007/s10661-024-12789-7

RESEARCH

IoT-based monitoring system and air quality prediction using machine learning for a healthy environment in Cameroon

Vitrice Ruben Folifack Signing · Jacob Mbarndouka Taamté · Michaux Kountchou Noube · Abba Hamadou Yerima · Joel Azzopardi · Yvette Flore Tchuente Siaka · Saïdou

Received: 20 December 2023 / Accepted: 6 June 2024 / Published online: 15 June 2024

Abstract This paper is aimed at developing an air quality monitoring system using machine learning (ML), Internet of Things (IoT), and other elements to predict the level of particulate matter and gases in the air based on the air quality index (AQI). It is an air quality assessor and therefore a means of achieving the Sustainable Development Goals (SDGs), in particular, SDG 3.9 (substantial reduction of the health impacts of hazardous substances) and SDG 11.6 (reduction of negative impacts on cities and populations). AQI quantifies and informs the public about air pollutants and their adverse effects on public health. The proposed air quality monitoring device is low-cost and operates in real-time. It consists of a hardware unit that detects various pollutants to assess air quality as well as other airborne particles such as carbon dioxide (CO_2), methane (CH_4), volatile organic compounds (VOCs), nitrogen dioxide (NO_2), carbon monoxide (CO), and particulate matter with an aerodynamic diameter of 2.5 microns or less ($PM_{2.5}$). To predict air quality, the device was deployed from November 1, 2022, to February 4, 2023, in certain bauxite-rich areas of Adamawa and certain volcanic sites in western Cameroon. Therefore, machine learning algorithm models, namely, multiple linear regression (MLR), support vector regression (SVR), random forest regression (RFR), XGBoost (XGB), and K-nearest neighbors (KNN) were applied to analyze the collected concentrations and predict the future state of air quality. The performance of these models was evaluated using mean absolute error (MAE), coefficient of determination (R-square), and root mean square error (RMSE). The obtained data in this study show that these pollutants are present in selected localities albeit to different extents. Moreover, the AQI values obtained range from 10 to 530, with a mean of 132.380 ± 63.705, corresponding to moderate air quality state but may induce an adverse effect on sensitive members of the population. This study revealed that XGB regression performed better in air quality forecasting with the highest R-squared (test score of 0.9991 and train score of 0.9999) and lowest RMSE (test score of 1.5748 and train score of 0. 0073) and MAE (test score of 0.0872 and train score of 0.0020), while the KNN model had the worst prediction (lowest R-squared and highest RMSE and MAE). This embryonic work is a prototype for projects in Cameroon as measurements are underway for a national spread over a longer period of time.

V. R. Folifack Signing · J. Mbarndouka Taamté · M. Kountchou Noube · A. Hamadou Yerima · Y. F. Tchuente Siaka (✉) · Saïdou
Research Centre for Nuclear Science and Technology, Institute of Geological and Mining Research, P.O. Box 4110, Yaoundé, Cameroon
e-mail: siakaf@yahoo.fr

J. Azzopardi
Department of Artificial Intelligence, Faculty of Information and Communication Technology, University of Malta, Msida, Malta

Saïdou
Nuclear Physics Laboratory, Faculty of Science, University of Yaoundé I, P.O. Box 812, Yaoundé, Cameroon

Keywords Internet of Things (IoT) · Air pollution · Air quality index (AQI) · Machine learning (ML) · Data analysis

Introduction

Air quality monitoring is of increasing concern today due to the development of urban and industrial areas. People suffer from health problems related to prolonged exposure to polluted environments (Budi et al., 2024; Fenger, 1999; Manisalidis et al, 2020; Molina & Molina, 2004). Yet, air is one of the most important natural resources for all life on planet Earth. All living organisms need good air quality, free from harmful gases and fine particles, to survive (Omer, 2008). Living beings, humans in particular, are largely responsible for atmospheric pollution through automobile emissions, industries, deforestation, and other anthropogenic emissions (Bisht et al., 2020; Li et al., 2020). Therefore, it is important to practice good behavior to preserve the environment and thus the health of the entire ecosystem. Good air quality is one of the SDGs recommended by the WHO, which can be expected in all cities around the world for the well-being of the population (Fund, 2015).

Air pollution issues have led organizations and individuals to become interested in air quality as it affects all members of various ecosystems. This requires the development of air quality management and control strategies to prevent diseases caused by prolonged exposure or poor air quality such as cardiopulmonary diseases, asthma, headaches, and dizziness (Banerjee & Srivastava, 2011; Zhu et al., 2017). Several pollutants are naturally and/or artificially responsible for air pollution, namely, nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), ozone (O_3), methane (CH_4), suspended particulate matter (particles with a diameter of 10 microns or less (PM_{10}), particles with a diameter of 2.5 microns or less ($PM_{2.5}$), smoke, and aerosols), carbon dioxide (CO_2), and volatile organic compounds (VOCs). These pollutants come from a variety of sources, including pollution from motor vehicles and aircraft, household smoke and heating or combustion appliances, waste incineration, and fires (Mishra & Goyal, 2015; Soni & Patel, 2017).

To effectively combat air pollution and protect public health, Kalaivani et al. performed air quality prediction and monitoring using an IoT sensor based on a machine learning algorithm (Kalaivani & Mayilvahanan, 2021). They described various ML algorithm studies related to air quality prediction and monitoring based on IoT sensor data. They also summarized historical and real-time data based on air quality forecasting model tools and techniques and described the merits and drawbacks of recent research methodologies. Martin-Baos and colleagues conducted IoT-based monitoring of air quality and road traffic using regression analysis (Martín-Baos et al., 2022). They compared the regression methods, namely, linear regression (LR), Gaussian process regression (GPR), and random forest (RF) using the performance measures R-square, MSE, MAE, and RMSE to significantly improve air quality index (AQI) estimates. Sonawani and Patil performed real-time air quality measurements, predictions, and alarms using a transfer learning-based IoT system for assisted living in an environment (Sonawani & Patil, 2024). They showed the significant importance of using low-cost sensors, with advanced open-source technology and improving the performance of the prediction capacity of new systems. Ansari and Alam developed a smart air pollution forecasting model based on IoT cloud and using univariate time series analysis (Ansari & Alam, 2024). They demonstrate how the proposed model produced significantly better results than its competitor, providing the most accurate and efficient forecasting model, with an MSE of 632.200, RMSE of 25.14, AE Med of 19, 11, a maximum error of 51.52, and an MAE of 20:49.

In air quality measurement, it is possible to consider the main pollutants in order to assess the overall air quality. An air quality indicator based on major pollutants, known as air quality index (AQI), allows the public to easily know whether air quality is good or poor (S. Kumari & Jain, 2018). The main pollutants defined by the World Health Organization (WHO) and other regulatory agencies used to calculate AQI are NO_2, SO_2, CO, O_3, PM_{10}, and $PM_{2.5}$ (Hu et al., 2015). In addition to indicating the level of danger, AQI informs emergency plans to be implemented in the event of a polluted index and can be used to alert the public to the dangers of air pollution (Kelly et al., 2012).

In this article, a smart real-time air quality monitoring device was locally realized and used to assess air pollutants such as CO_2, CH_4, VOCs, NO_2, and $PM_{2.5}$ in the Adamawa Bauxite and Western Volcanic

regions of Cameroon. The objective is to analyze the relationship between weather factors or anthropogenic activities and AQI values and investigate how machine learning models can be applied to predict AQI values based on air pollution data. Due to demographic expansion, the growing number of vehicles and factories pollute the air in major cities at an alarming rate (Pucher et al., 2007). Therefore, some models of machine learning algorithms such as multiple linear regression (MLR), support vector regression (SVR), random forest regression (RFR), XGBoost (XGB), and K-nearest neighbors (KNN) were applied to analyze the changes in concentrations of measured pollutants in order to predict the future state of air quality in both regions of the country. Machine learning is an important tool for predictive analytics, and its potential benefits are significant (Sun & Scanlon, 2019). The performance of these models was evaluated using mean absolute error (MAE), root mean square error (MSE), coefficient of determination (R-square), and root mean square error (RMSE).

The remainder of this article is organized as follows: "Related surveys" reviews some related work, "Materials and methods" presents the materials used to implement the intelligent air quality monitoring system are proposed. The AQI evaluation methods, air quality prediction models, and some system performance analysis methods are also presented within this section. The results of exploratory data analysis, AQI assessment, prediction, and comparison are reported in "Results and discussion." The conclusion is presented in "Conclusion."

Related surveys

In an era marked by significant strides in industrial and transportation sectors, coupled with the exponential surge in urbanization and human demographics, the symbiosis between humans and their immediate environment has never been more apparent. The fabric of daily existence grows increasingly interwoven with the environmental tapestry, rendering human health and activities particularly susceptible to any prospective environmental degradation (Manisalidis et al., 2020). Within this dynamic, air quality prediction emerges as a task of daunting complexity, due to the capricious characteristics and significant temporal and spatial heterogeneity of airborne contaminants and particulates (Usmani, 2022).

As a result, the ability to develop low-cost air quality monitoring devices has increased, particularly in urban areas where the negative impacts of air pollution on public health and the biosphere are most evident. As globalization ebbs, digitalization is gaining importance and heralding the emergence of Artificial Intelligence (AI) as a cornerstone of the contemporary industrial revolution. AI and machine learning (ML) stand at the forefront of this evolution, poised to enhance living standards by optimizing outcomes through augmented efficiency, celerity, and precision in both mundane and developmental undertakings (Pradeep et al., 2018).

There are numerous attempts in the academic and scientific fields to use ML algorithms to forecast air quality. A notable instance of such innovative exertions emanates from Cambridge, UK, where a holistic methodology was adopted, amalgamating diverse variables such as pollutant concentration, urban traffic, aerial imagery, and weather conditions to feed an array of predictive models, namely, statistical analysis, machine learning, and deep learning utilizing neural networks (Babu Saheer et al., 2022). These models span the spectrum from traditional statistical evaluation to cutting-edge machine learning and deep learning, the latter exploiting the capabilities of neural networks. In models using Weather Normalized Models (WNM). Performance appraisals of such ML approaches integrated within Weather Normalized Models (WNMs) unveil the superior efficacy of neural networks and deep learning techniques to predict pollutants (Chau et al., 2022).

Further diversifying the landscape, support vector machine methodology was used for forecasting AQI, demonstrating commendable accuracy, substantiated by a robust coefficient of determination (R-square), and minimal mean square deviations (Leong et al., 2020). Contrastingly, Taiwanese studies utilized ML to forecast particulate matter (PM) concentrations, outstripping the traditional models in performance, as corroborated by multiple performance metrics including R-square, RMSE, MAE, and MSE (Harishkumar et al., 2020).

Expanding the methodological arsenal, logistic regression, classification arbors, and random forests have been employed in tandem with expert-driven and automated variable selection techniques to predict

PM_{10} levels in Bogotá, Colombia. Interestingly, variables handpicked by specialists outperformed those chosen through automated mechanisms (Martínez et al., 2018). In a similar vein, the air quality in Delhi, India, was accurately predicted hourly using eXtreme Gradient Boosting (XGB) and random forest (RF) algorithms (Juarez & Petersen, 2021). Meanwhile, a comparative study in Beijing revealed gradient boosting (GB) as superior to XGB in accuracy, forecasting prowess, and operational efficiency (Su, 2020). Furthermore, a hybrid approach combining RF and SVM was effectively used in Rio de Janeiro, Brazil, to explore the nexus between various air pollutants and meteorological variables, yielding precise ozone level predictions (de Oliveira et al., 2021).

In a practical application of these principles, Sai et al. introduced an air quality monitoring contraption employing MQ135 and MQ7 sensors (Sai et al., 2019). This apparatus, linked to IoT platforms like ThingSpeak and Cayenne, aims to heighten public awareness of environmental perils. Additionally, the data collated were subjected to ML algorithm analysis to decipher the embedded information. Presently, meteorological forecasting hubs are pivoting towards AI-optimized observations, computations, and analyses to not only achieve enhanced outcomes but also expedite forecasts while curtailing computational demands. Through meticulously tailored algorithms, AI has the potential to refine the preliminary processing of atmospheric and physical data utilized in weather forecasting models.

In Africa, most work using ML algorithms is focused on fine and coarse particulate matter prediction, particularly in South African cities, Uganda, and sub-Saharan Africa, to name a few (Adong, Bainomugisha et al., 2022; Coker et al., 2021; Morapedi & Obagbuwa, 2023; Zhang et al., 2021). To the best of the author's knowledge, no work has been done in Cameroon to predict air quality, but ML algorithms have been used to predict land cover at different geographical scales in the North, daily direct solar energy, spatial prediction of metal traces and pollution indices in sediments of a mining site, and flood prediction in certain regions, to name a few (Dtissibe et al., 2024; Yaulande et al., 2022; Yuh et al., 2023).

In light of the above, a prediction of air quality based on a set of data from a locally developed electronic device is investigated. This study delineates the deployment of an economical air quality monitoring mechanism that can detect a range of pollutants including CO_2, CH_4, VOCs, NO_2, CO, and $PM_{2.5}$. Our focus was trained on the bauxite-rich Adamawa region and certain volcanic locales in Cameroon's Western region. The collected data were harnessed for forecasting models, followed by an exhaustive analysis using statistical and relevant ML models to predict pollution based on AQI, resulting in an optimized model with near-perfect accuracy after careful performance metrics evaluation.

Materials and methods

Smart air quality device

The realized real-time device is based on the Atmega 2560 microcontroller, a board designed to support more complex projects. It is an ideal board for 3D printing and robotics projects. The system also includes a power supply, an LCD display, a micro-SD card module, and several sensors, including a methane sensor (MQ3), a carbon monoxide and nitrogen dioxide sensor (CJMCU-6814), an equivalent carbon dioxide (eCO2) sensor, and a total volatile organic compound (TVOC or VOC) sensor (CCS811), as well as a dust sensor (PPD42NS). It consists of electronic sensors in contact with atmospheric pollutants, a central processing unit which is the Atmega 2560 microcontroller board, and many output organs. Among these organs, there is XBee wireless transmission module for IoT technology application. A schematic diagram (Fig. 1a) and the final realization of the smart device (Fig. 1b) are shown.

The XBee S2C modules use the ZigBee radio communication protocol based on the IEEE 802.15.4 standard with an operating frequency of 2.4 GHz (Bhavanam & Ragam, 2023). They enable remote data transmission between the air quality device and a computer via two XBee transmitter and receiver modules and an XCTU application previously installed on the PC. These modules are used for the IoT components in this article. The wireless data transmission scheme between air quality device and remote PC performed in this work is shown in Fig. 2.

During the operation of the air quality measurement device, the IoT system ensures the real-time transmission and reception of data on the remote-control PC whose screen zoom is shown in Fig. 3.

Low-cost electronic components

The gas sensors respond to toxic gases such as carbon monoxide (CO), nitrogen dioxide (NO_2), equivalent carbon dioxide (eCO_2), particulate matter ($PM_{2.5}$), total volatile organic compounds (TVOC), and methane (CH_4). The MQ3 sensor is highly sensitive to methane and has a long lifetime. The CJMCU-6814 consists of three sensor chips with independent heaters and sensitive layers. It is used here to detect NO_2 and CO detection. The PPD42NS sensor provides reliable data on $PM_{2.5}$; particulate matter of 2.5 µm size in mg/m^3 is converted to µg/m^3. XBee S2C modules enable remote data transmission between the air quality monitor and a computer via XCTU application. The technical characteristics of the electronic sensors used are listed in Table 1, as indicated in the ref (Jacob et al., 2021).

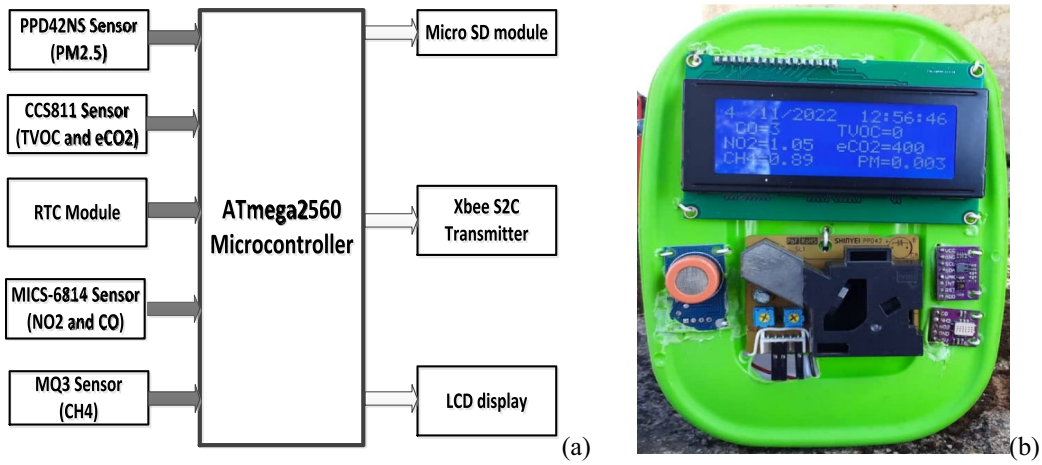

Fig. 1 (**a**) schematic diagram and (**b**) the final realization of air quality device

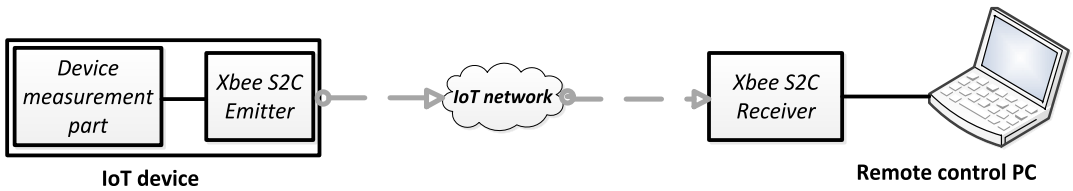

Fig. 2 Schematic diagram of wireless transmission between air quality device and remote PC controller

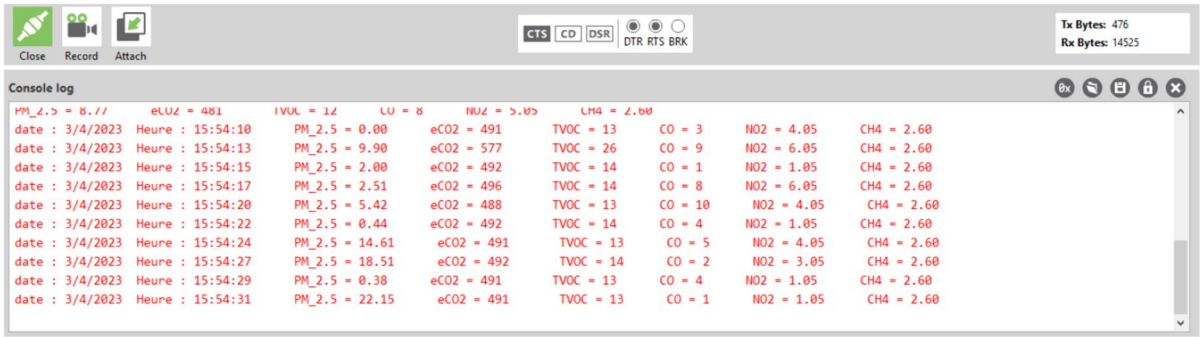

Fig. 3 Air quality data reception on remote PC controller

Sensor's configuration and calibration

Sensor calibration and IoT module configuration were performed prior to the commencement of any. Each sensor was calibrated with parameters according to the manufacturer's procedure. The calibration of the MQ3 sensor was performed according to the pre-calibration algorithm used in (Idrees et al., 2018; Jacob et al., 2021). Calibration of the CJMCU-6814 sensor (NO_2 and CO) is performed using the IDE calibration code provided by the manufacturer to obtain the complete configuration of the sensor (Sensortech, 2015). Recalibration is then carried out in an air-conditioned room and takes half an hour. The PPD42NS dust sensor was calibrated according to the procedure described in (Jacob et al., 2021; Venkatraman Jagatha et al., 2021). The basic procedure for calculating the suspended particle concentration was also used in the Integrated Development Environment (IDE) program. The classical equation proposed in (Alam et al., 2017), which gives the concentrations of particulate matter (in µg/m^3), has been proven to work as follows:

$$PM_{2.5\_}Concentration = 0.62+520.0*Ratio - 3.8*Ratio^2+1.1*Ratio^3 \quad (1)$$

where

$$Ratio=\frac{\sum low\ pulse\ width}{sample\ period} \quad (2)$$

Experimental calibration and comparison of measurements were performed using a simultaneous measurement with the GENT sampler (reference device for measuring suspended particulate matter (Nducol et al., 2021a, b) and PM sensors at the same site and at the same point (Fig. 4). Both instruments measure $PM_{2.5}$ concentrations. The ratio of the concentration obtained by the two instruments provides a calibration factor for configuring the electronic sensors.

A statistical analysis based on linear regression is used, and the value of the correlation coefficient R considered determines the level of appreciation of the electronic method in relation to the GENT Sampler.

The results of the data obtained are shown in Fig. 5, which illustrates the $PM_{2.5}$ concentrations for each method used. The correlation factor is used to correct the discrepancy in order to obtain a good agreement between the two methods (Fig. 6).

Equation (3) defines the relationship between the two methods:

$$\text{Electronic PM} = 1.8779 * \text{GENT PM} + 2.0995 \approx 2 * \text{GENT PM} \quad (3)$$

This shows that the electronic device is in agreement with the GENT Sampler if the relationship (4), derived from Eq. (3), is respected:

$$(\text{Electronic PM} - \text{GENT PM})/\text{Gent PM} \approx 1 \quad (4)$$

The verification of Eq. (4) was done by averaging the concentrations of the different $PM_{2.5}$ listed in Table 2.

By applying the values obtained in Eq. (4), we obtain a reliability coefficient (1.00 ± 0.01) of the device produced in relation to the GENT.

The same procedure was used for the other sensors used in this work.

Air quality index and exposure assessment methods

Air quality assessment requires two approaches, namely, air monitoring, which is either direct

Table 1 Technical characteristics of the used electronic sensors

Types of sensors	Pollutants and gases measured	Detection capability	Sensor tolerance (%)	Operating energy (V_{CC})	Preheat time (hours)
PPD42NS	PM1.0 (mg/m^3)	0 ~ 1.4	± 5	4.5 ~ 5.5	≤ 24
	PM2.5 (mg/m^3)		± 10		≤ 24
MQ3	CH_4 (ppm)	0.01 ~ 2	± 3	5 ± 0.1	≤ 24
CCS811	CO_2 (ppm)	400 ~ 64,000	± 3	5 ± 0,1	≤ 24
	COV (ppb)	0 ~ 64,000	± 3		≤ 24
CJMCU-6814	NO_2 (ppm)	0.05 ~ 10	± 2	4.9 ~ 5.1	≤ 24
	CO (ppm)	1 ~ 1000	± 1.2		≤ 24

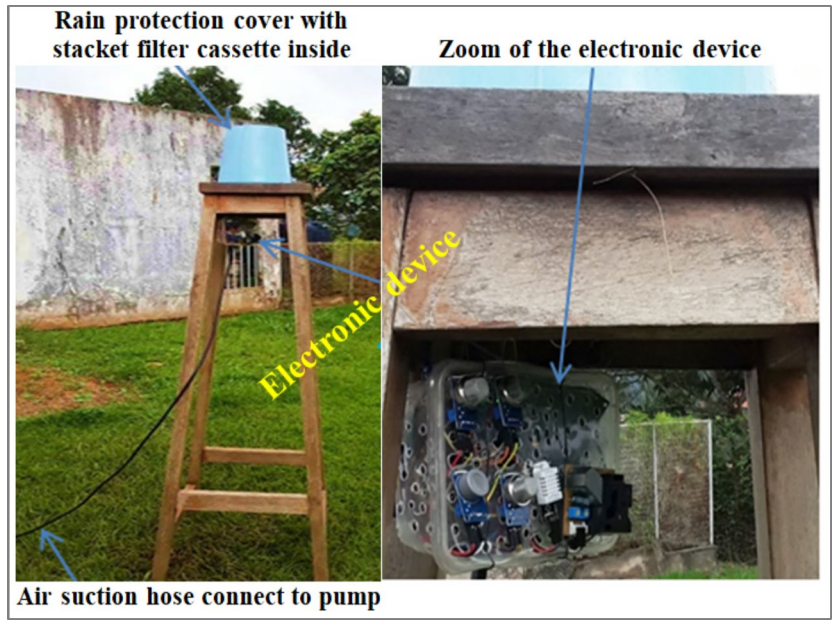

Fig. 4 The both sampling system as assembled in the field (Nducol et al., 2021a, b)

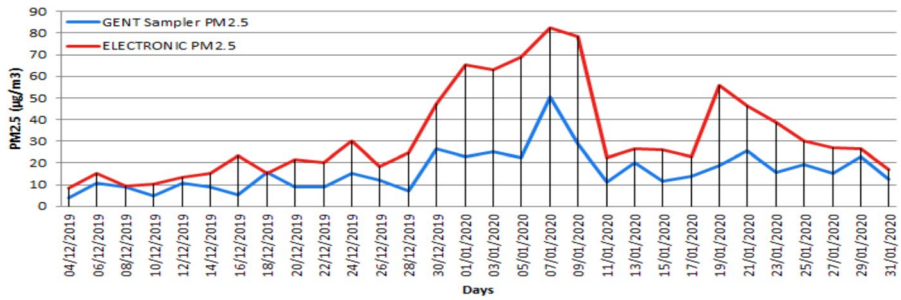

Fig. 5 $PM_{2.5}$ concentration of GENT Sampler and proposed device (Jacob et al., 2021)

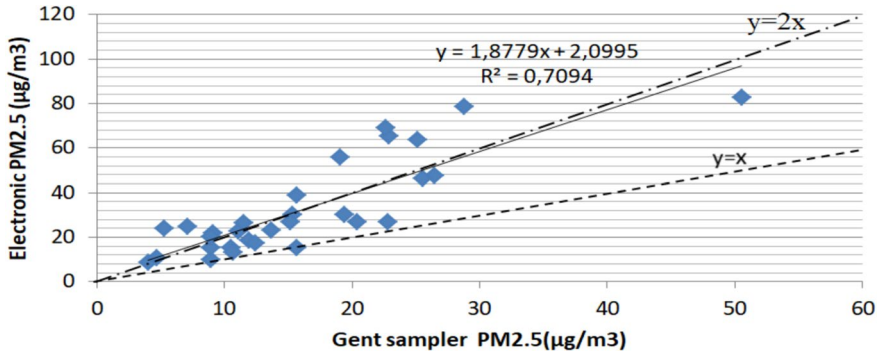

Fig. 6 Correlation between concentration of the GENT Sampler and electronic device (Jacob et al., 2021)

measurements (in situ monitors) or indirect (fixed monitors) and biological measurements, which use biological markers to assess human exposure. The work related to air monitoring techniques is based on direct measurements with locally made monitors. Therefore, it is necessary to analyze and interpret the data obtained and compare them with the recommended thresholds. The recommended thresholds for

Table 2 Average values of GENT and electronic device

Measuring period		
Device	GENT $PM_{2.5}$	Electronic $PM_{2.5}$
Average concentration ($\mu g/m^3$)	16.117 ± 0.11	32.367 ± 0.40

the measured pollutants are those of WHO (World Health Organization, 2010), as shown in Table 3. These values are used to calculate the air quality sub-index according to Eq. (5), and the final AQI is the highest value of all sub-indices.

$$\text{IQA sub} - \text{index} = (\text{measure}/\text{reference value}) \times 50 \quad (5)$$

AQI is a powerful indicator that quantifies air quality (Kumari et al., 2020). It is a number used to show the public how polluted the air is or is likely to become. It is possible that as the AQI increases, more and more of the population will be affected and this will have a significant adverse impact on health. Different countries have their own air quality indices to meet different national air quality requirements (Kumari et al., 2020; Yu et al., 2016). The choice of AQI is linked to the maximum of different sub-AQIs given by Eq. (6).

$$\text{AQI} = \text{Max}\{\text{AQI}_1, \text{AQI}_2, \text{AQI}_3, \ldots, \text{AQI}_n\} \quad (6)$$

According to the WHO guidelines, recommendations, and reference values applied in this work, three pollutants ($PM_{2.5}$, CO, and NO_2) were allowed to obtain the AQI values. Figure 7 explains the operating diagram of the AQI system based on Eq. (6).

For this work, the AQI is evaluated from the pollutants $PM_{2.5}$, CO, and NO_2 which determine α_1, α_2, and α_3 from the reference values of each pollutant (Table 3). An obtained single number (index value) corresponds to the nomenclature and color for different pollutants given in Table 4. In this table, there are six AQI categories, namely, Good, Satisfactory, Moderate, Poor, Very Poor, and Severe which define the potential health effects, corresponding to a color code (Alsaedi & Liyakathunisa, 2019).

Table 3 Pollutants responsible for air quality and their WHO reference values (World Health Organization, 2010)

Pollutants	WHO guideline values	Health effects
Particles with a diameter less than 2,5 µm ($PM_{2.5}$)	15 $\mu g/m^3$ daily average 5 $\mu g/m^3$ annual average	Risks of developing cardiovascular and respiratory diseases
Ozone (O_3)	100 $\mu g/m^3$ on average over 8 h	May induce breathing difficulties, asthma. Risk of disruption of the functioning of the lungs
Nitrogen dioxide (NO_2)	10 $\mu g/m^3$ annual average 25 $\mu g/m^3$ hourly average	Risk of development of chronic bronchitis in asthmatic subjects
Sulfur dioxide (SO_2)	40 $\mu g/m^3$ daily average	May cause respiratory and pulmonary system function and eye irritation
Carbon monoxide (CO)	60 $\mu g/m^3$ daily average	Asphyxiant gas that attaches to red blood cells and prevents them from carrying oxygen properly in the body

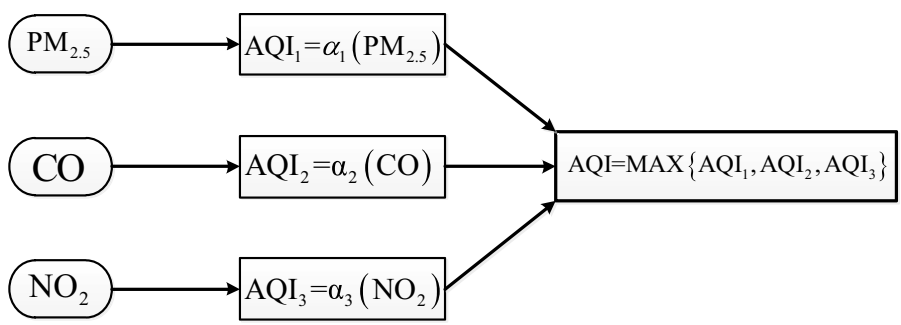

Fig. 7 Block diagram for assessing the AQI value

Data collection and preprocessing

Geolocation of measurement sites

The present work is carried out in two regions of Cameroon (Adamawa Region and West Region) with different geological characteristics. The Adamawa region is in a bauxite zone, and the West region is in a volcanic zone. The study area in the West region is chosen for this study because of its geological type (volcanic origin) and population density (Momo et al., 2020). Similarly, the study area of the Adamawa region is selected based on its geological type (bauxite). These study areas have been geolocated and shown in yellow in Fig. 8 consisting of two divisions of the Western Region and two divisions of the Adamawa Region.

The air quality monitoring system used in this work, consisting of low-cost gas and suspended particulate matter sensors, was deployed in five districts (subdivisions) in the selected areas of Fig. 8, which are shown in Fig. 9. A measurement campaign of each target area of Fig. 9 was carried out in the Minim and Martap subdivisions of the Adamawa Region, Bandjoun, Baham, and Bayangam of Western Region as indicated on the map of Fig. 9. For each monitoring site, the concentrations of $PM_{2.5}$, CO_2, VOC, CH_4, CO, and NO_2 are collected and stored in the SD media.

Since this study deals with the monitoring and prediction of outdoor air quality, 6 parameters described were collected during the daytime, including traffic and peak hours, from 8 a.m. to 5 p.m. over a period from November 1, 2022, to February 4, 2023. Cameroon has two climatic seasons, a dry season from November to March and a rainy season from April to October. The measurement period considered in this work corresponds to the dry season, with its potential for high levels of certain pollutants. These monitoring sites constitute a sample that will serve as a reference for such an initiative in Cameroon.

Table 4 Features of AQI and color indicator

AQI level	Category	Health concern	Color
0–50	Good	Good	Green
51–100	Satisfactory	Moderate	Yellow
101–150	Moderately polluted	Unhealthy for sensitive groups	Orange
151–200		Unhealthy	Red
201–300	Poor	Very unhealthy	Purple
301–400	Very poor	Hazardous	Maroon
401–higher	Severe		

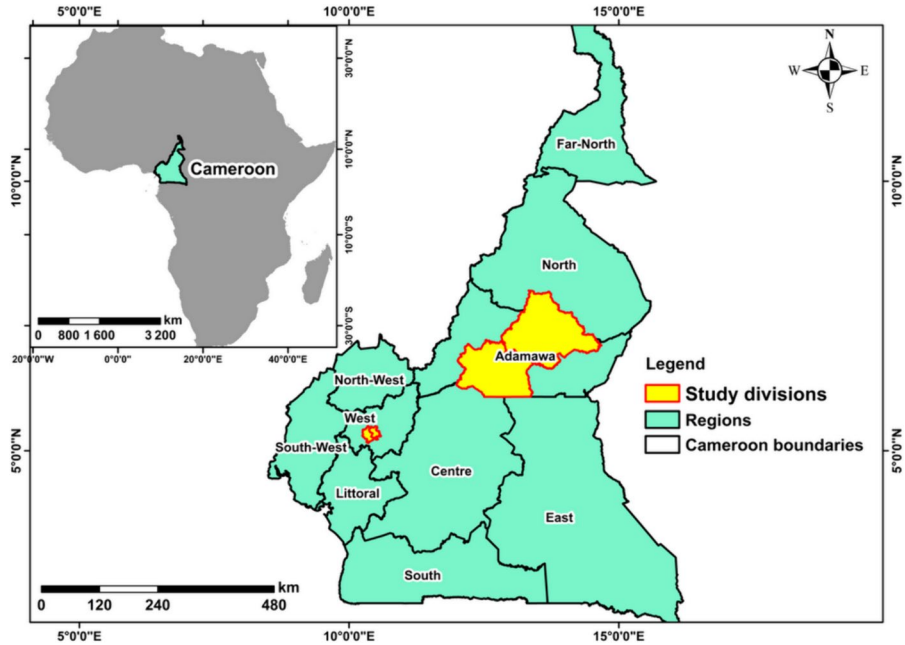

Fig. 8 Geolocation of target departments for air quality monitoring

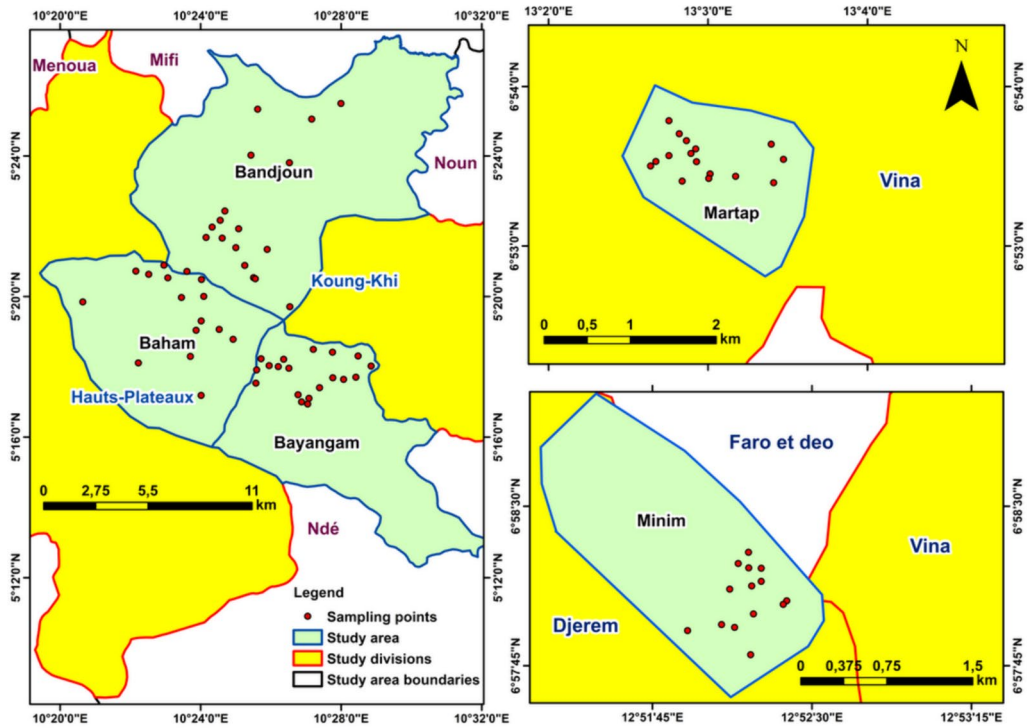

Fig. 9 Highlighting the different measurement positions (samples) in the different study areas

Each examined area in Figs. 8 and 9 differ based on their geological structure. Table 5 shows the parameters of all collected data by locality, type of geology, number of measured parameters, and measurement period. This geological structure influence naturally some pollutants without forgetting the human activities of the areas.

The study of lakes or volcanic and geothermal areas in the West Cameroon region shows that they emit several atmospheric pollutants, in particular, gases (Edmonds et al., 2018). Among these gases, significant amounts of carbon dioxide CO_2 (Nyos and Monoun, in Cameroon) and methane CH_4 (Lake Kivu in Rwanda) are generally emitted (Sigurdsson, 1988). In some volcanic lakes, gases such as sulfur dioxide (SO_2), hydrogen sulfide (H2S), carbon monoxide (CO), hydrogen halides (hydrogen chloride HCl and hydrogen fluoride HF), and radon emitted during the volcanic eruption may still be present in the air in trace amounts (Jourdain et al., 2016). Radon is a radioactive gas produced by the decay of uranium from deep within the Earth. In volcanic lakes, radon is emitted from the ground along with CO_2. In its mining process, bauxite is crushed and then mixed with soda at high temperature and under pressure. The obtained liquor, sodium aluminate, is stripped

Table 5 Dataset description

Locality	Geology	Number of measured parameters	Duration
Minim Martap	Bauxite and metamorphic	6	November 1, 2022 to February 4, 2023
Bandjoun Baham Bayangam	Volcanic and metamorphic		

of impurities and then diluted and cooled, resulting in the precipitation of hydrated alumina oxide. The mining and improper transportation of this mineral often lead to soil erosion, deforestation, air, and water pollution, with severe impacts on agriculture and fisheries (Ray & Ray, 2011). A study of air pollution in bauxite areas shows the presence of several gases and particles, including $PM_{2.5}$, CO_2, VOC, CH_4, CO, and NO_2 (Adak & Kour, 2021). Bauxite in the Adamawa region will be explored in the near future, and protecting the population from pollution is essential. This work in the bauxite zone is part of a preliminary study, carried out to know the level of air pollution before the bauxite mining works.

Data preprocessing

Data preprocessing is a critical step in any data analysis where the collected data are transformed into a structured form that can be used for ML models. Unclean data impact forecasting process and can lead to poor predictions. Raw sensor data are collected without any filtering process (Sharma et al., 2021). The filtering process is implemented in a different environment (storage system) rather than on the embedded device to reduce the complexity of the embedded system. Common data errors in real-time monitoring equipment and devices are outliers and missing data. Data may contain unwanted values (nulls, voids, and symbols) on some rows and variables. Data preprocessing obeys certain criteria before being introduced into ML algorithms, namely, removing not-a-number (NaN) data, filling NaN data to zero, or interpolating data (Goh et al., 2021; Ni et al., 2009).

The device has been programmed to record data from each parameter every 3 s, which corresponds to 20 data per minute and 1200 data per hour for optimal operation. The original dataset consisted of 12,453 rows and 9 columns including date, time, locality, and the six measured parameters. The undesirable data in this dataset are null (zero), empty/NaN, and outliers.

Handling missing or null values Finding any null values and replacing or removing them, depending on the dataset, is the most important part of the preprocessing steps (Halsana, 2020). Since the air quality monitor used in this work was fed discretely (from one measurement area to another) from one point to another. Some sensors, such as the CCS811 sensor (for CO_2 and VOC detection) and the PPD42NS sensor (for $PM_{2.5}$ detection), were found to have a latency period before measuring normally. During this waiting period, the data are not collected correctly and it is necessary to remove this data during this period of operation. The data preprocessing operation in this work involved deleting of some specific rows. Lines consisting of zero values and NaN (corresponding to the latency phase of some sensors when switching the device on and/or off). These values were removed for the simple reason that they cannot be used. After this step, the dataset is reduced to 11,865 rows.

Handling outliers An outlier is a value that differs significantly from the overall trend of other observations when considering a dataset with common characteristics. It is very important to check for outliers in the dataset and apply an imputation method. When outliers are managed, the dataset becomes normally distributed, which is a useful characteristic for a dataset to improve prediction accuracy (Goh et al., 2021; Halsana, 2020). In this study, data interpolation is performed using the nearest neighbor method up to order ten for distant positions. The nearest known neighbor value is used to replace the outlier. This method uses nine different formula states depending on the position of the outlier. The outlier position, located from the tenth position, is replaced by the average of the ten previous data. This interpolation is summarized by Eq. 7, where x is the outlier at position i.

$$x_i = \frac{1}{i}\sum_{n=1}^{i} x_{i-n} \tag{7}$$

Machine learning approaches

The forecast performed in this work is a set of processes from data collection to the prediction of the air quality level based on the AQI. Figure 10 shows the different steps involved in the air quality prediction process. This flowchart in Fig. 10 is grouped into several steps including data collection, exploratory data analysis, definition of the ML model, training of the model, and analysis of the results. Thus, the collected data must be cleaned in the preprocessing step and then explored to get a general idea of the data behavior. In this step, the data must be observed through

various statistical descriptive analyses and processing tools. The shape of the data allows us to know if the data are uniform or if there are outliers, which is very important in the prediction process. In the learning process, the processed dataset is divided into two subsets of data, namely, the training set and the test set to form the model.

In general, the training set represents 60–80% of the dataset and the test set represents 20–40%. Empirical studies have shown that the best results are obtained by dividing the dataset in an 80:20 ratio (Gupta et al., 2023). In this process, random sampling is used to split the data into training and test sets. Thus, 80% of the data is used for training and 20% for test sets. After this split, the training data can be used to train and predict the model.

The experiments conducted in this work were performed on an Inter(R) Core(TM) i7-3630QM CPU @ 2.4 GHz, 4 cores, (8 CPUs), and 12 GB of RAM, running the Microsoft Windows 11 Professional operating system. The ML models were built in the Jupyter Notebook 6.4.8 environment, using the Python 3.6 programming language. The Panda, scikit learn, and matplotlib libraries were used to perform preprocessing and data exploratory, ML algorithms, and graphics respectively. Sklearn metrics were also used for performance evaluation.

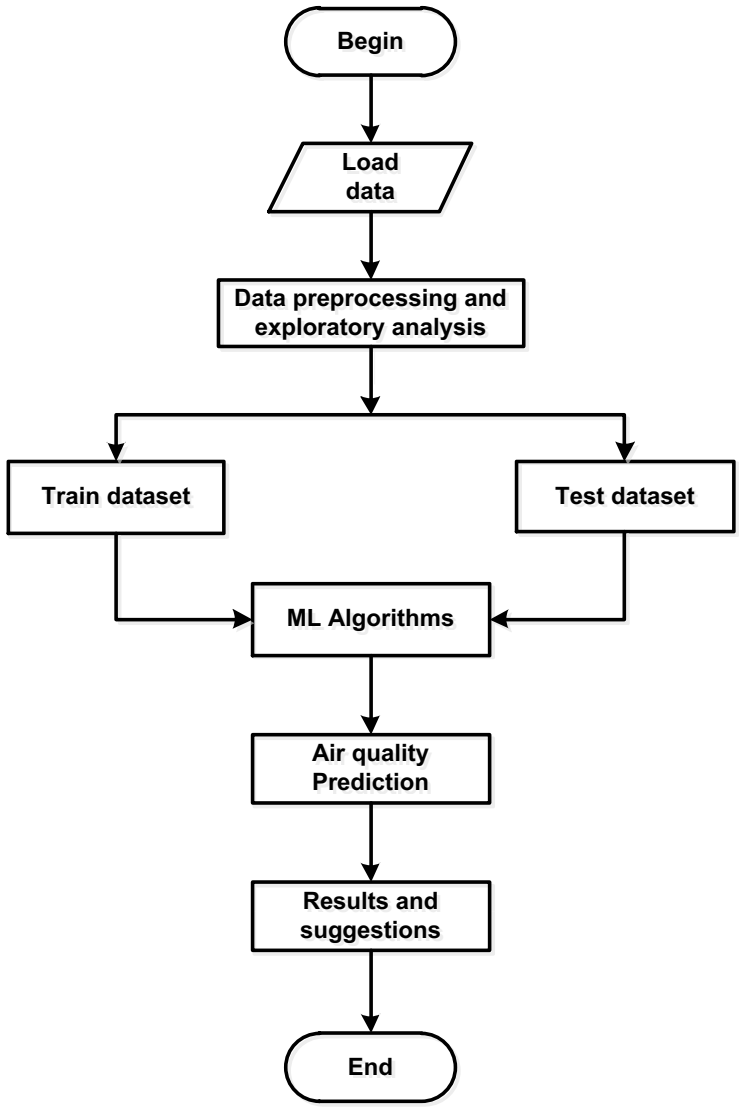

Fig. 10 Process flowchart and processing

There are a large number of ML algorithms that could be applied to air quality forecasting tasks. The choice of algorithm depends on the objective, the type of input data, the resources available, the size of the dataset, and many others. It is not easy to compare these algorithms, since a particular algorithm may show different results in different circumstances. All it takes is for the input data to change, for example, and the algorithms would require different parameters to work properly (Fiandrino et al., 2020).

This work considers linear ML models (MLR and SVR) and ensembles (RF, XGB, and KNN). Linear models are interpretable and fast to train. Linear models cannot recover nonlinear dependencies without data preprocessing. Ensemble models, on the other hand, have been used to eliminate the possibility of missing complex data correlations. Quality can be increased considerably with the appropriate tuning of their hyperparameters. However, it should be remembered that the selection of hyperparameters for ensembles is time-consuming.

Multiple linear regression (MLR)

MLR is a statistical technique commonly used to predict the outcome of a given set of data based on the value of two or more variables. In other words, MLR examines how multiple independent variables relate to one dependent variable. MLR creates a linear relationship that approximates the best or majority of individual data points. MLR forecasts the future of a dependent variable (target) by establishing linear relationships with one or more independent variables (predictors). The prediction is based on the assumption that the relationship between the target and the predictors is dependent or causal.

The advantages of MLR include its simplicity and the ability to have all potential variables in a single model (Lei et al., 2019; Lei et al., 2020; Lei et al., 2022). MLR predicts by calculating a weighted sum of the elements of the input, plus a constant (bias term), as in Eq. (8).

$$y_i = \alpha_o + \alpha_1 x_1 + \alpha_2 x_2 \cdots + \alpha_n x_n + \delta \tag{8}$$

where α_o is y-intercept (constant term), y_i dependent variables, x_i explanatory variables, $\alpha_i (i \neq o)$ slope coefficients for each x_i, and δ the model's error term (the residuals).

Support vector regression (SVR)

Support vector regression (SVR) is a machine learning technique that uses supervised learning and constructs hyperplanes to separate different classes. SVR is typically used to analyze data with a categorical output variable. This technique can support applications with linear or nonlinear data in predictive processes. The general idea in SVR is to find the hyperplane that represents the minimum distance between itself and the data points. SVR is well suited for nonlinear data, as it uses a so-called kernel function to map nonlinear data to a higher-dimensional space, in which the model can then fit a linear hyperplane. The advantages of SVR include its robustness to outliers, high prediction accuracy, and ease of implementation (Lei et al., 2022; Rybarczyk & Zalakeviciute, 2018).

Finding the best line is the main principle of SVR. Thus, the hyperplane with the most points is the one that processes the data best. However, the first idea is always to minimize the error by individualizing the hyperplane that maximizes the margin, taking into account that in most cases, some error is tolerated.

Random forest regression (RF)

A random forest regression is a meta-estimator that fits a set of classification decision trees to different subsamples of the dataset, using the mean to improve prediction accuracy and control overfitting. It ensures that each tree in the set is generated from a sample of the training set (Breiman, 2001). Also, during tree generation, the selected partition is the best one in a random subset of features, rather than the best partition among all alternatives.

As a result, each tree of the random forest is trained on a random subset of data according to the bagging principle, with a random subset of features according to the random projections principle. Since the data are quantitative, the predictions are then averaged. As a consequence of this randomness, the bias of the forest may increase slightly, but due to the mean, its variance is generally reduced, which can compensate for the increased bias, resulting in an overall superior model.

XGBoost (XGB)

XGBoost (or eXtreme Gradient Boosting) is an improved model of the Gradient Boost algorithm based

on sequential ensemble learning and decision trees. This ML algorithm is also used to solve common business problems with minimal resources. The XGB algorithm is currently considered the best in its class. It is particularly useful when dealing with small or medium-sized structured or tabular data (Chen et al., 2015). It can thus be used to solve regression problems, but it can also solve categorization or classification problems.

XGB is a machine learning technique that builds strong learners from weak ones. First, an initial model is built from the data, and then a second model is created with the goal of correctly predicting observations that the first model failed to predict. The combination of these two models is expected to be better than either model alone. This boosting process is then repeated several times, with each new model attempting to improve on the errors made by the previous models.

The uniqueness of this algorithm lies in the decision tree used. In addition to being fast, accurate, and efficient, XGB offers several advantages, including bias reduction, ease of implementation, and very efficient computation.

K-nearest neighbors (KNN)

KNN algorithm is a supervised learning method. It can be used for both regression and classification. In general, the KNN algorithm relies on the entire dataset. For a value that is not part of the dataset to be predicted, the KNN technique will search for the K instances that are closest to the observation. Thus, the algorithm will use their output variables to estimate the value of the observation variable to be predicted (Abu Alfeilat et al., 2019; Kramer & Kramer, 2013). For regressions, KNN considers the mean (or median) of the variables from the K closest observations for prediction. When KNN is used for classification, the mode of the variables of the K-nearest observations is used for prediction. In this work, this technique is used for regression problems.

During the training phase, the KNN algorithm stores all training data for reference. When making predictions, the algorithm searches for the K instances in the dataset that are closest to the observation by calculating the distance between the input data point and all training examples. Then, for these K neighbors, the algorithm uses their output variables to calculate the value of the observation variable we want to predict.

System performance evaluation

The evaluation of an analysis and prediction model is performed using performance metrics. In this work, root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R-square), and precision are used.

Root mean square error and mean absolute error

Root mean square error (RMSE) is calculated based on the mean squared error (MSE), which is the standard error measure for evaluating the performance of regression models. This tool is useful for evaluating performance during training and for defining a cost function due to its simplicity (Goh et al., 2021). If there is one statistic that normally takes precedence over the others, it is RMSE, which is defined as the square root of the mean square error defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2} \tag{9}$$

where m is the number of samples, $\hat{y}_i$ is the predicted value, and y_i is the actual value of a sample. Another error measure used to assess the performance of regression models is the mean absolute error (MAE), defined as:

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}\left|y_i - \hat{y}_i\right| \tag{10}$$

Geometrically, RMSE is interpreted as the average distance of the points from the regression model. It weights errors in proportion to their size, whereas MAE weights all errors equally. For this reason, RMSE is more sensitive to outliers than MAE. Thus, MAE is considered a better metric for measuring average performance, while RMSE is better for measuring a model's sensitivity to outliers. It is then recommended to use these two sets of metrics as they provide complementary information for evaluating the performance of a model (Horton, 1940).

Coefficient of determination and accuracy

The coefficient of determination, or R-square, is used when looking for the strength of fit between

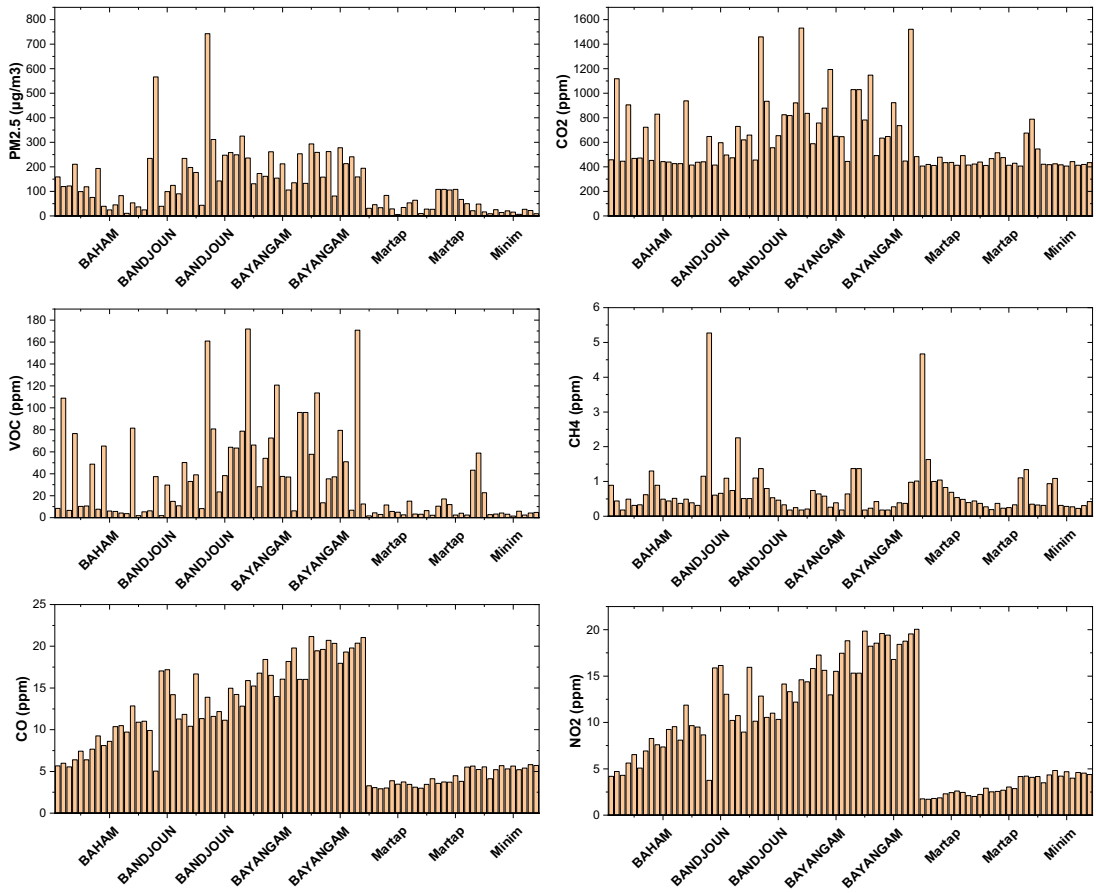

Fig. 11 Overall distribution of raw values for the different pollutants obtained by location

Table 6 Descriptive statistical study of the hold dataset

Statistical parameter	Locality	$PM_{2.5}$ (µg/m3)	CO_2 (ppm)	VOC (ppm)	CH_4 (ppm)	CO (ppm)	NO_2 (ppm)
Mean	-	128.851	621.313	33.634	0.691	10.292	9.215
SE of mean	-	13.362	29.133	4.420	0.085	0.643	0.651
Minimum	Martap	5.360	406.282	1.487	0.180	2.910	1.717
25%	Baham	34.518	432.769	4.722	0.310	5.228	4.142
50%	Bandjoun	105.665	480.592	12.153	0.450	9.810	8.461
75%	Bayangam	194.958	741.986	51.670	0.841	15.924	14.774
Maximum	Bayangam	742.280	1531.307	171.920	5.270	21.174	20.054

a regression model and the data collected. Specifically, the coefficient of determination is an index of the quality of a regression's prediction. The coefficient of determination is between 0 and 1. The closer it is to 1, the better the regression fits the collected data. One is equal to 100%, in which case, the result can be predicted without error from the independent variables. Conversely, if the index is close to zero, it indicates the near absence of correlated data and the result cannot be predicted by

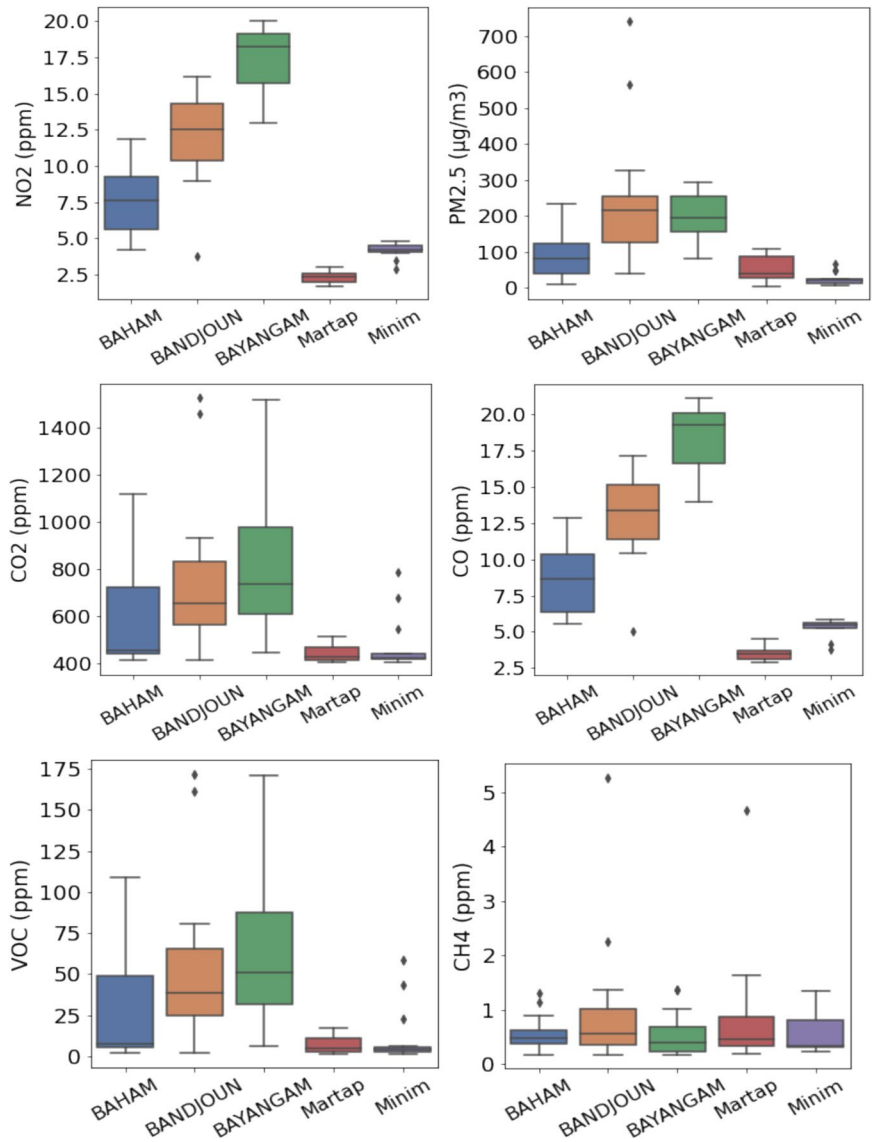

Fig. 12 Box plot diagrams showing the distribution of significant values obtained from the dataset

any of the independent variables. The coefficient of determination is useful for predicting future events based on the probability provided by the result of its calculation. Therefore, it is necessary to have as much data as possible so that the result is as accurate as possible (Chicco et al., 2021).

Its evaluation is based on the quality of the data among the totality of the recorded data (TSS); we count those that constitute only residual variants (RSS). The following calculation allows us to find a more suitable and accurate coefficient of determination.

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \tag{11}$$

Results and discussion

Exploratory data analysis (EDA)

Using statistical graphs and other data visualization techniques, exploratory data analysis (EDA) is a method for thoroughly analyzing and scrutinizing datasets in order to highlight their key features. EDA provides a better understanding of the variables in the dataset and the relationships between them. It can also help determine whether the statistical techniques being considered for data analysis are appropriate. A general observation of each measured parameter is shown by the histograms in Fig. 11 and according to

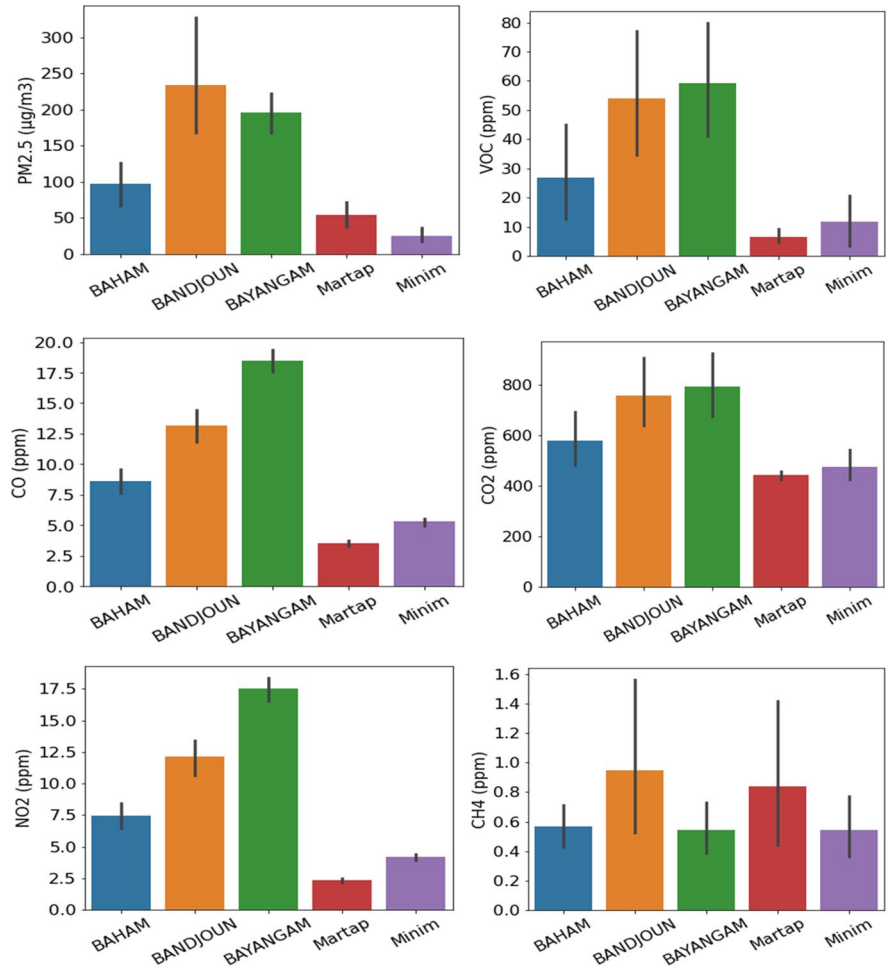

Fig. 13 Histogram of mean concentration of the different pollutants obtained according to the different localities

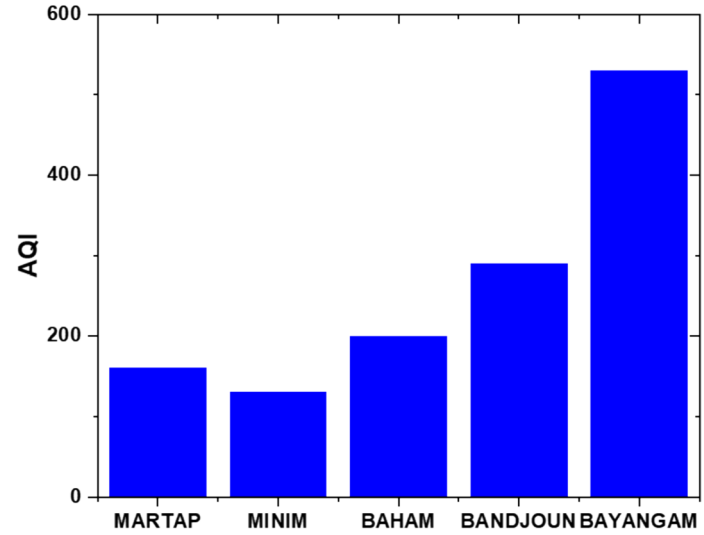

Fig. 14 Bar chart distribution of maximum AQI values obtained by locality

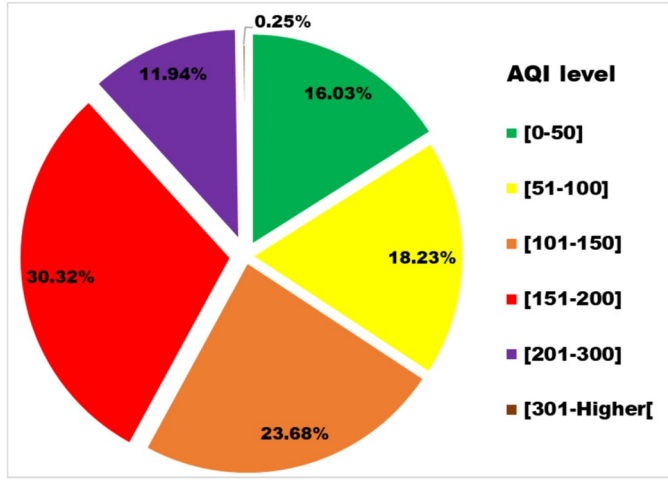

Fig. 15 Distribution of the air quality index by pollution level and percentage

Table 7 Average AQI values of each locality and corresponding categories

Locality	Mean AQI	AQI category
Martap	36.07 ± 065	Good
Minim	54.65 ± 0.83	Satisfactory
Baham	104.08 ± 0.60	Moderate
Bandjoun	136.15 ± 0.62	Moderate
Bayangam	186.03 ± 1.01	Moderate

the different localities. A high concentration can be seen in the Bandjoun and Bayangam localities.

In addition, the Minim and Martap sites have a low concentration of all measured gases, except CO_2. For a better interpretation of the obtained concentrations, an in-depth statistical study is made below. It is important to mention that the presentation of Fig. 11 allows to identify extreme values through the peaks (max and min). The cleaning of the dataset is done in the preprocessing for a good prediction. An overview of descriptive statistics is presented in Table 6 and shows the mean, standard deviation, minimum, and representative values (minimum, first quartile, second quartile (median), third quartile, and maximum).

A detailed representation of each measured parameter can be seen in the box plot of Fig. 12 according to the different localities. This diagram visually summarizes a variable, identifies the extreme values, and helps understand the concentrations distribution. The whiskers on either side of the box represent the bottom of 25% and the top of 25% of the data value ranges, excluding outliers. The center line of each colored box is the median, corresponding to half of the observations. The rectangular box itself is the interquartile range (75–25%), which represents half (50%) of the data.

The different statistical indicators (maximum, minimum, and median) and bands showing element dispersion are provided by the box plot in Fig. 12. In these diagrams, outliers are observed on all parameters. As a result, outliers are observed in Banjoun and Mimim localities for the majority of pollutants except CH_4. It is important to mention that these outliers are handled as specified in the preprocessing process for better prediction. For a discrete representation of the different mean values of all measured pollutants, the corresponding histograms by locality are shown in Fig. 13. High values are observed in Bandjoun and Bayangam localities, except for $PM_{2.5}$ and CH_4. This high representation highlights the possible sources of different pollutants in these localities.

Air quality index

The AQI values obtained in this work range from 10 to 530, with an average of 132.380 ± 63.705. The set of AQI by localities is represented by the histogram of Fig. 14. From this figure, it can be seen that the highest AQI value is in Bayangam and the lowest in Martap. This observation provides a general overview of the level of air quality based on the calculation of AQI. Based on this, it is clear that the air in the Bayangam region is more polluted than in the Martap region, where the air quality is good and safe.

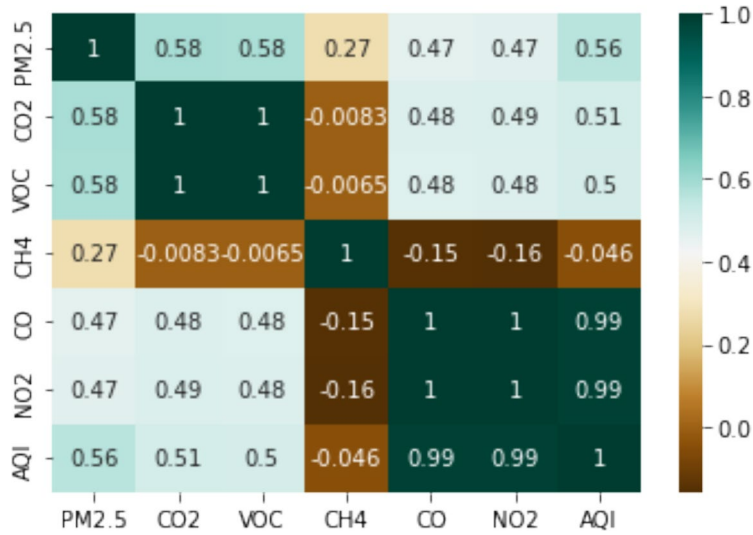

Fig. 16 Correlation matrix between pollutants and AQI

Table 8 Characteristic elements of the bauxite zone dataset

Label	Features
Dataset shape	(2343, 9)
Locality	Minim and Martap
AQI range	[10–103]
AQI category	Good, Satisfactory, and Moderate
Training set shape	(1874, 5)
Test set shape	(1874, 1)

For correct interpretation according to the color code linked to Table 4, a breakdown of the AQI per value range and corresponding color is shown in Fig. 15. Similarly, the percentage attribution to each AQI range obtained is also shown. Figure 15 is a quantitative and qualitative representation of the AQI obtained in this work. In light of this, all the intervals from Table 4 are shown in Fig. 15. The categories of Good, Satisfactory, and Moderately polluted represent almost half in terms of percentage of the dataset, i.e., 16.03% for safe beaches, 18.23% for satisfactory data, and 11.49% for moderate. On the other hand, the majority of data obtained, i.e., more than 50%, are considered polluted and dangerous. Based on these analyses, in-depth studies should be carried out in the affected areas and other surrounding areas, as well as strict behavioral measures for the population to avoid prolonged exposure and a high health risk.

It is important to have an idea of the overall air quality at different measurement sites. AQI helps to report and categorize air quality, indicating how clean or polluted the air is and the potential health effects associated with it. The higher the AQI, the more polluted the air is and the greater the health problem. For the present work, the average pollution level is evaluated using the average IQA for each locality and presented in Table 7. This table shows good air quality for Martap, satisfactory for Minim, and moderate air quality for Baham, Bandjoun, and Bayangam.

Correlation analysis

Correlation testing is an important part of model training. The goal of this analysis is to evaluate the correlation of each parameter with the target. In fact, it is advisable to remove features that correlate poorly with the target, as this could lead to poor predictions (Ameer et al., 2019; Halsana, 2020). A high correlation between features and the target leads to a good prediction performance. This analysis is done using the correlation matrix shown in Fig. 16, which shows the correlation between all pollutants and AQI.

This figure is very important for forecasting purposes because it is selective and allows an overall statement to be made about the relationship with the target variable. In essence, a strong correlation should be observed between the target value, the AQI, and its components, namely, CO, $PM_{2.5}$, and NO_2. In fact, there is an almost perfect correlation of 0.99 between AQI, CO, and NO_2, meaning that the latter makes a major contribution to

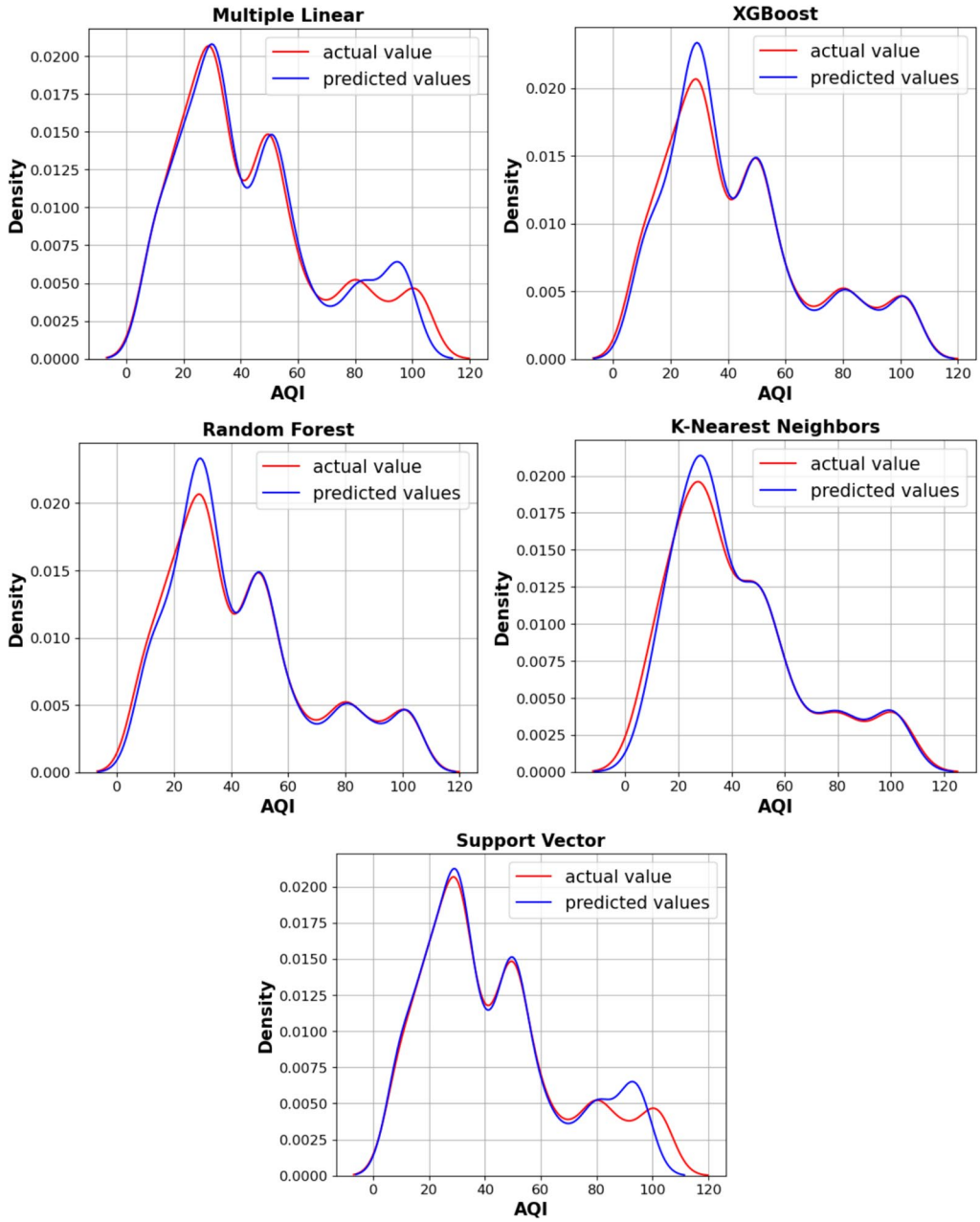

Fig. 17 Distribution of predicted and actual AQI values according to the different ML models used in the bauxite zone

achieving AQI. Although $PM_{2.5}$ is an integral part of the calculation of AQI, its contribution to its prediction is only moderate, with a correlation of 0.56. It is also important to mention that other pollutants (CO_2, VOC, and CH_4) can also influence the predictions as they are an integral part of the dataset. From the correlation matrix in Fig. 16, it is clear that CO2 and VOC have an average correlation with AQI comparable to the contribution of PM2.5, and therefore, they can have a positive impact on AQI level prediction.

The predictive relevance of attributes is low if their correlation with the target variable (IQA) is

Table 9 Comparison of regression techniques according to performance metrics of bauxite regions

Algorithm			Performance indicator			
	Test R-square	Train R-square	Test RMSE	Train RMSE	Test MAE	Train MAE
KNN	0.9879	0.9514	2.7844	5.6755	1.0483	1.4923
SVR	0.9613	0.9641	5.0135	4.9921	1.2045	1.1408
XGB	0.9634	0.9707	4.8781	4.5045	1.2049	1.1632
MLR	0.9627	0.9658	4.9239	4.8676	1.6485	1.5819
RF	0.9634	0.9707	4.8729	4.5047	1.2131	1.1626

Table 10 Characteristic elements of the volcanic zone dataset

Label	Features
Dataset shape	(9522, 9)
Locality	Baham, Bayangam, and Bandjoun
AQI range	[20–295]
AQI category	Good, Satisfactory, Moderate, and Poor
Training set shape	(7617, 5)
Test set shape	(7617, 1)

zero or close to zero. Consequently, it would be crucial to remove all variables that significantly and negatively affect the prediction's quality. As a result, CH_4 is removed from the training set due to its weak and negative correlation with AQI. The features data (training) include CO_2, VOC, CO, $PM_{2.5}$, and NO_2, and the test set is that of the target variable (AQI). The quantity (number of rows) depends strongly on the parameters of the "train_test_split" function of the ScikitLearn library (sklearn). With an 80:20 separation ratio, i.e., 20% for test set and 80% for training set, the training set and test set have 9492 rows.

Forecasting, evaluation, and comparisons

The data were gathered from various locations situated in two distinct geographical zones, namely, the bauxite zone and the volcanic one. This section begins with analyses in each region, followed by a global analysis to determine which model is optimal.

Bauxite area

After preprocessing the data from this area, the dataset consists of the following elements, which are summarized in Table 8.

Using the different ML techniques, the results of the analysis are shown in Fig. 17. This figure shows the density distribution plots of the actual (red curve) and predicted (blue curve) AQI. Almost all models have a mismatch between actual and predicted values. For example, only SVR has a good fit between the two curves for low values and a poor fit like MLR for high AQI values. RF and XGB have a similar prediction pattern like KNN, but the latter has a better fit, which can be seen in the good agreement between predicted and actual values. This observation fully supports the hypothesis that KNN is more suitable for small datasets (Ching et al., 2022).

For a more detailed analysis of these trends, the values of the test set and training set performance metrics have been calculated and presented in Table 9. In light of this table, KNN shows the best test performance, i.e., the highest R-squared and lowest RMSE and MAE, and the worst training set performance, i.e., the lowest R-squared and the highest RMSE and MAE. From this observation, we cannot conclude whether KNN is the best model. At the same time, the other models show a remarkable similarity between the test set and train set. In particular, the training scores are in some way an improvement over the test scores. In fact, further analysis with a large amount of data is likely to lead to satisfactory and more credible results.

Volcanic area

The data in this zone are more important than that of the bauxite zone. The characteristics of this data are described in Table 10 where all of which are superior to those of the bauxitic zone dataset. This claims better performance of the techniques used and adapted to the large dataset.

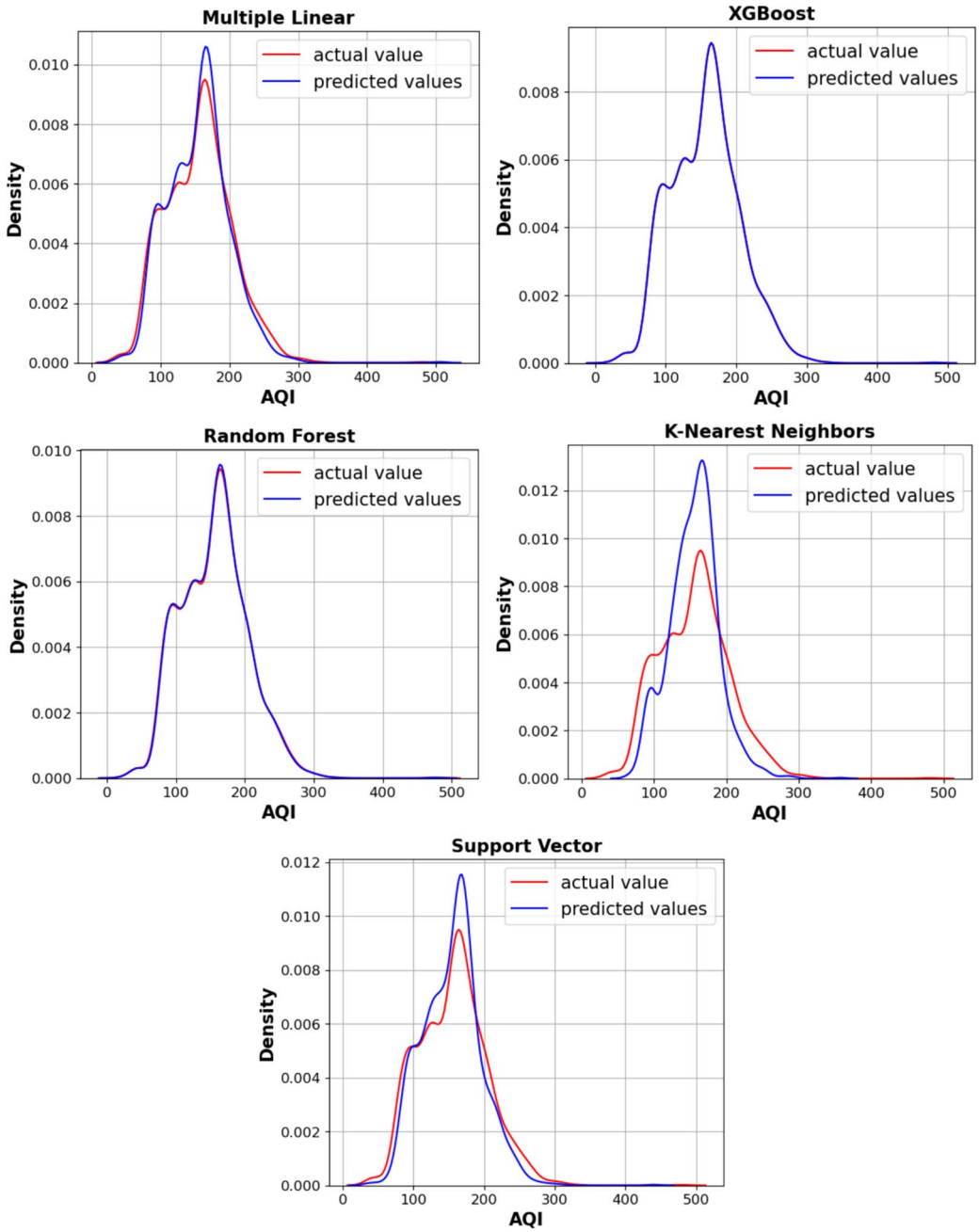

Fig. 18 Distribution of predicted and actual AQI values according to the different ML models used in the volcanic zone

The evolution of the predicted AQI densities of each algorithm is shown in Fig. 18, where the blue curve shows the evolution of the predicted data and the red one the evolution of the actual data. From this figure, it can be seen that XGB and RF are the best because the two curves (actual and predicted) are merged (for XGB) or almost (for RF). MLR gives an acceptable prediction, but SVR and KNN

Table 11 Comparison of regression techniques according to performance metrics of volcanic regions

Algorithm			Performance indicator			
	Test R-square	Train R-square	Test RMSE	Train RMSE	Test MAE	Train MAE
KNN	0.7545	0.7759	23.6905	23.5812	16.4236	16.1619
SVR	0.8684	0.9200	17.3435	14.0820	11.3589	9.1526
XGB	0.9999	0.9999	0.3997	0.0097	0.0231	0.0038
MLR	0.8907	0.9684	15.8053	8.8419	10.3603	5.4271
RF	0.9924	0.9953	1.1480	1.5034	0.7968	0.8493

Table 12 Characteristic elements of the dataset of all study areas

Label	Features
Dataset shape	(11,865, 9)
Locality	Minim, MartapBaham, Bayangam, and Bandjoun
AQI range	[10–315]
AQI category	Good, Satisfactory, Moderate, Poor, and Very poor
Training set shape	(9492, 5)
Test set shape	(9492, 1)

give poor predictions. Contrary to the conclusion reached in Fig. 17, KNN shows a poor prediction in Fig. 18. This result was expected, as it confirms the hypothesis that KNN is a poor candidate for predicting large datasets (Ching et al., 2022).

A quantitative analysis is performed using performance indicators and presented in Table 11. In this table, performance indicators for test set and train set were determined. It can be seen that XGB has better performance in terms of R-squared, RMSE, and MAE. A constant evolution is observed between the test data and train data indicators, with improved performance for train data.

Overall measured data

To ensure the best XGB performance, all data from the bauxitic and volcanic zones were combined. The characteristics of the common dataset are shown in Table 12. We can see the impact of the outlier management on the maximum AQI value (315 instead of 530). This treatment includes five categories ranging from good to very poor.

Figure 19 shows the actual (red curve) and predicted (blue curve) IQA density plots obtained from the different prediction models. The good performance of XGB and RF can be seen and confirmed by the similarity between predicted and actual values. MLR shows an intermediate performance, while SVR and KNN perform poorly, with a visible gap between the two curves.

The performance indicators for training set and test set were calculated and shown in Table 13. These indicators show that XBG has the best performance on both test set and train set. From these results, we can safely conclude that XGB is the best prediction model for the studied dataset. It can also be seen that KNN with the least R-squared and large RMSE and MAE values shows poor prediction.

The correlation results between the actual and predicted AQI data for each model considered are shown in Fig. 20. On the graphs in Fig. 20, the regression line for each model is plotted between the predicted data and the original data. From this graph, we can see that the XGB model has a high accuracy, as indicated by the R-squared previously evaluated in Table 13.

These results tell us how accurate our predictions are and how far they deviate from actual values. Since higher R-squared values indicate a better fit between model predictions and actual observations, XGB and RF are the better models, while KNN is the poor model. Based on this information, we can classify the ML models used as follows: KNN < SVR < MLR < RF < XGB.

Hyperparameter-tuning

Hyperparameters directly control model structure, function, and performance. By adjusting hyperparameters, model performance can be fine-tuned for

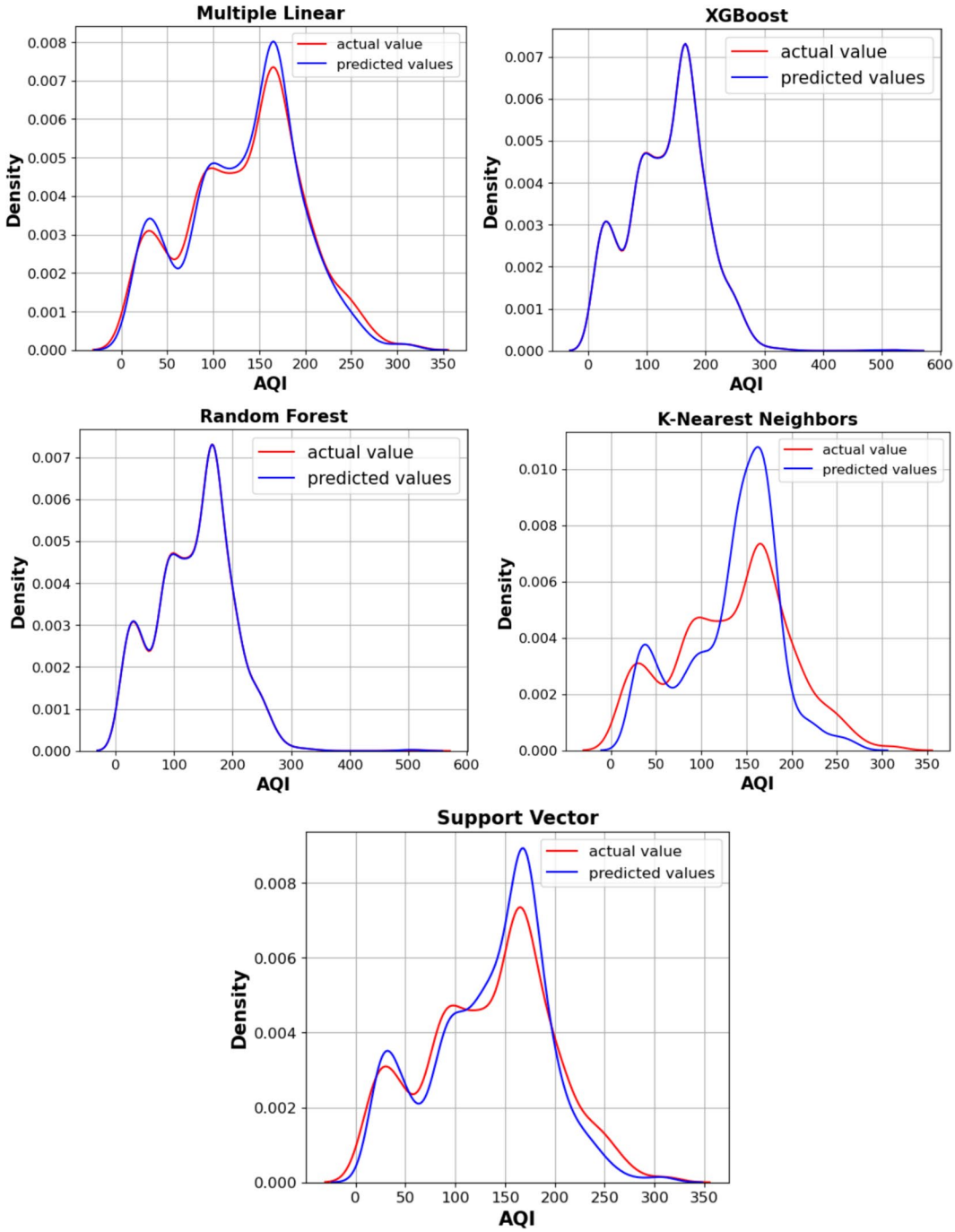

Fig. 19 Distribution of predicted and actual AQI values according to the different ML models used in the entire zone

Table 13 Comparison of regression techniques according to performance metrics of all regions

Algorithm			Performance indicator			
	Test R-square	Train R-square	Test RMSE	Train RMSE	Test MAE	Train MAE
KNN	0.8387	0.8464	25.6150	25.1406	17.2925	16.9162
SVR	0.9351	0.9492	16.2502	14.3100	10.6236	9.3784
XGB	0.9991	0.9999	1.5748	0.0073	0.0872	0.0020
MLR	0.9585	0.9886	12.9792	6.7656	8.0055	3.8518
RF	0.9987	0.9990	2.2727	0.5357	0.1473	0.0387

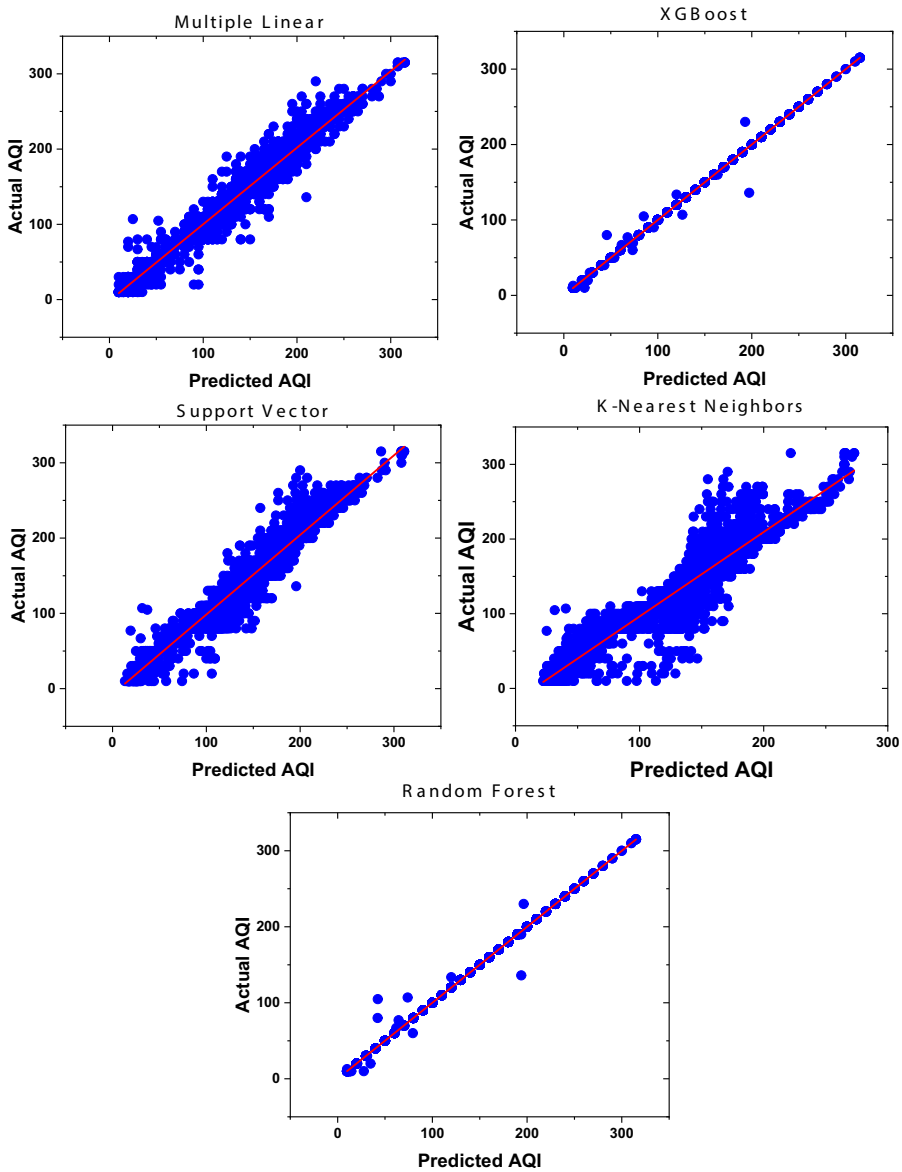

Fig. 20 Correlation between predicted and actual AQI for each model used

optimal results. This process is an essential part of machine learning, and choosing the right hyperparameter values is crucial to success. There are many algorithms for tuning hyperparameters, but the most commonly used types are Bayesian optimization, grid search (used in this analysis), and random search.

With this algorithm, a list of hyperparameters and a performance measure is specified, and the algorithm examines all possible combinations to determine the best fit. Grid search works well but is relatively tedious and computationally demanding, especially with a large number of hyperparameters. An overfitted model may look impressive on the training set but will be useless in a real application. Therefore, the standard hyperparameter optimization procedure accounts for overfitting through cross-validation. Hyperparameter analysis is performed on the best estimated models, RF and XGB, to be sure of the parameters that make them efficient.

Hyperparameters in XGB In XGB, there are tree-specific and learning task-specific hyperparameters. The tree-specific hyperparameters used are:

- max_depth (maximum tree depth): Deeper trees have the potential to overfit data, but they can also capture more complex patterns.
- min_child_weight (minimum sum of instance weights required): This can be used to control tree complexity by preventing the creation of leaves that are too small.
- Subsample (percentage of rows used for each tree construction): Reducing this value can prevent overlearning by training on a smaller subset of data.

The hyperparameter specific to the learning task is eta (also known as learning rate): reduction in step size used in updates to avoid overfitting. Lower values make the model more robust by taking smaller steps.

The grid of search parameters considered for this evaluation is defined in Table 14.

After training, the optimal hyperparameters are learning_rate = 0.3, max_depth = 3, min_child_weight = 2, n_estimators = 100, and subsample = 1. With these hyperparameters, the improved R square scores are 0.9998 for training set and 0.9992 for test set. These scores obtained after application of the best hyperparameters are not very far from those obtained previously; this further justifies the performance of this model.

Hyperparameters in RF For this model, the set of tuning hyperparameters are as follows: number of trees in the forest (n_estimators), maximum number of features considered for splitting a node (max_features), maximum number of levels in each tree (max_depth), minimum number of data points placed in a node before the node is split (min_samples_split), minimum number of data points allowed in a leaf node (min_samples_leaf), and method for sampling data points (bootstrap) with or without replacement.

The grid of search parameters considered for this evaluation is defined in Tables 14 and 15

The best hyperparameters are bootstrap = True, max_depth = 20, max_features = 3, min_samples_leaf = 1, min_samples_split = 2, and n_estimators = 200 for a test score of 0.9993 and a training score of 0.9999. These scores obtained after application of the best hyperparameters are not very far from those obtained previously; this further justifies the performance of this model.

Table 14 Search parameter grid for XGB

Hyperparameters	Set of values
max_depth	[3, 4, 5]
learning_rate	[0.01, 0.1, 0.2, 0.3]
n_estimators	[100, 200, 250, 300]
subsample	[0.5, 0.7, 1]
min_child_weight	[2, 3, 4]

Table 15 Search parameter grid for XGB

Hyperparameters	Set of values
max_depth	[5, 10, 15, 20]
max_features	[2, 3]
n_estimators	[100, 200, 300]
min_samples_leaf	[1, 2, 4]
min_samples_split	[2, 5, 10]
bootstrap	[True]

Conclusion

In this paper, an air quality monitoring system using Internet of Things (IoT) technology was developed. Artificial Intelligence/machine learning techniques were explored to predict the level of particulate matter and gases in the air based on the air quality index (AQI). The air quality monitoring device used in this work is low-cost and operates in real-time. This device consists of a hardware unit (low-cost sensors) that detects various pollutants (CO_2, CH_4, VOC, NO_2, CO, and $PM_{2.5}$), a processing unit (Arduino programmable board), and a wireless communication system (XBee modules). To predict air quality levels, the device was used from November 1, 2022, to February 4, 2023, in the selected volcanic area of the West Region and the bauxite area of the Adamawa Region of Cameroon. ML regression models, namely, multiple linear regression (MLR), support vector regression (SVR), random forest regression (RFR), XGBoost (XGB), and K-nearest neighbors (KNN) were used to analyze and predict the collected concentrations. The performance of these models was evaluated using the mean absolute error (MAE), coefficient of determination (R-squared), and root mean square error (RMSE). This study highlighted the importance of having a good dataset, free of impurities, and, above all, with a large amount of data. As a result, KNN regression is more suitable for small datasets, while XGB regression performs better for large datasets. Thus, the highest correlation between predicted and actual data with the highest R-squared (test score of 0.9991 and train score of 0.9999) and lowest RMSE (test score of 1.5748 and train score of 0.0073) and MAE (test score of 0.0872 and train score of 0.0020) were obtained by XGB, while the KNN model had the worst prediction (lowest R-squared and highest RMSE and MAE). Tuning hyperparameters using grid search is not only specialized; the algorithm examines all possible combinations to determine the best fit and therefore requires significant runtime, especially with a large number of hyperparameters. Future work will focus on improving hyperparameter tuning for real-world applications while simultaneously improving the set of models used in comparisons.

Author contribution Vitrice Ruben Folifack Signing: data collection, conceptualization, analysis, investigation, validation, writing the main manuscript text; Jacob Mbarndouka Taamté: development of electronic devices, data collection, data processing and writing of articles published in air quality; Michaux Kountchou Noubé: data collection, data processing and writing of articles; Abba Hamadou Yerima: technical operator, data collection and data processing; Joel Azzopardi, Yvette Flore Tchuente Siaka, and Saïdou: supervision, read and approved the final manuscript.

Funding This work was supported by the Cameroon Public Investment Budget (BIP 2021-2022).

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

Ethical approval All authors have read, understood, and have complied as applicable with the statement on "Ethical responsibilities of Authors" as found in the Instructions for Authors.

References

Abu Alfeilat, H. A., Hassanat, A. B., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., & Prasath, V. S. (2019). Effects of distance measure choice on k-nearest neighbor classifier performance: A review. *Big data, 7*(4), 221-248. https://www.liebertpub.com/doi/abs/10.1089/big.2018.0175

Adak, P., & Kour, N. (2021). A review on the effects of environmental factors on plants tolerance to air pollution. *Journal of Environmental Treatment Techniques, 9*(4), 839–848. https://www.dormaj.org/index.php/jett/article/view/371

Adong, P., Bainomugisha, E., Okure, D., & Sserunjogi, R. (2022). Applying machine learning for large scale field calibration of low-cost PM2. 5 and PM10 air pollution sensors. *Applied AI Letters, 3*(3), e76. https://onlinelibrary.wiley.com/doi/full/10.1002/ail2.76

Alam, M., Khan, M. D., Khairulalam, M., Syed, A., Rajkumar, R., & Azam, T. B. (2017). Industrial level analysis of air quality and sound limits monitoring in Bangladesh using real time control system. *Vibroengineering Procedia, 16*, 81–86. https://www.extrica.com/article/19329

Alsaedi, A. S., & Liyakathunisa, L. (2019). *Spatial and temporal data analysis with deep learning for air quality prediction*. In 2019 12th International Conference on Developments in eSystems Engineering (DeSE) (pp. 581–587). IEEE. https://ieeexplore.ieee.org/abstract/document/9073002

Ameer, S., Shah, M. A., Khan, A., Song, H., Maple, C., Islam, S. U., & Asghar, M. N. (2019). Comparative analysis of machine learning techniques for predicting air quality in smart cities. *IEEE access, 7*, 128325–128338. https://ieeexplore.ieee.org/abstract/document/8746201/

Ansari, M., & Alam, M. (2024). An intelligent IoT-cloud-based air pollution forecasting model using univariate time-series analysis. *Arabian Journal for Science and Engineering, 49*(3), 3135-3162. https://link.springer.com/article/10.1007/s13369-023-07876-9

Babu Saheer, L., Bhasy, A., Maktabdar, M., & Zarrin, J. (2022). Data-driven framework for understanding and predicting air quality in urban areas. *Frontiers in Big Data,*

5, 822573. https://www.frontiersin.org/articles/10.3389/fdata.2022.822573/full

Banerjee, T., & Srivastava, R. K. (2011). Assessment of the ambient air quality at the Integrated Industrial Estate-Pantnagar through the air quality index (AQI) and exceedence factor (EF). *Asia-Pacific Journal of Chemical Engineering, 6*(1), 64-70. https://onlinelibrary.wiley.com/doi/abs/10.1002/apj.450

Bhavanam, B. P. R., & Ragam, P. (2023, December). Assessing the performance of ZigBee RF protocol using path loss models for IoT application. In International e-Conference on Advances in Computer Engineering and Communication Systems (ICACECS 2023) (pp. 348–359). *Atlantis Press*. https://www.atlantis-press.com/proceedings/icacecs-23/125995740

Bisht, A., Kamboj, N., Kamboj, V., & Bisht, A. (2020). A review on the role of emerging anthropogenic activities in environmental degradation and emphasis on their mitigation. *Archives of Agriculture and Environmental Science, 5*(3), 419–425. https://doi.org/10.26832/24566632.2020.0503025

Breiman, L. (2001). Random forests. *Machine learning, 45*, 5–32. https://doi.org/10.1023/a:1010933404324

Budi, H. S., Catalan Opulencia, M. J., Afra, A., Abdelbasset, W. K., Abdullaev, D., Majdi, A., Masoume, T., Hafez, A. E., & Mohammadi, M. J. (2024). Source, toxicity and carcinogenic health risk assessment of heavy metals. *Reviews on Environmental Health, 39*(1), 77–90. https://doi.org/10.1515/reveh-2022-0096

Chau, P. N., Zalakeviciute, R., Thomas, I., & Rybarczyk, Y. (2022). Deep learning approach for assessing air quality during COVID-19 lockdown in Quito. *Frontiers in Big Data, 5*, 842455. https://www.frontiersin.org/articles/10.3389/fdata.2022.842455/full

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). Xgboost: Extreme gradient boosting. *R Package Version 0.4–2, 1*(4), 1–4. https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peer J Computer Science, 7*, e623. https://peerj.com/articles/cs-623/

Ching, P. M. L., Zou, X., Wu, D., So, R. H. Y., & Chen, G. H. (2022). Development of a wide-range soft sensor for predicting wastewater BOD5 using an eXtreme gradient boosting (XGBoost) machine. *Environmental Research, 210*, 112953. https://www.sciencedirect.com/science/article/pii/S0013935122002808

Coker, E. S., Amegah, A. K., Mwebaze, E., Ssematimba, J., & Bainomugisha, E. (2021). A land use regression model using machine learning and locally developed low cost particulate matter sensors in Uganda. *Environmental Research, 199*, 111352. https://www.sciencedirect.com/science/article/pii/S0013935121006460

de Oliveira, R. C., Cunha, C. L., Tôrres, A. R., & Corrêa, S. M. (2021). Forecasts of tropospheric ozone in the Metropolitan Area of Rio de Janeiro based on missing data imputation and multivariate calibration techniques. *Environmental Monitoring and Assessment, 193*, 1-16. https://link.springer.com/article/10.1007/s10661-021-09333-2

Dtissibe, F. Y., Ari, A. A. A., Abboubakar, H., Njoya, A. N., Mohamadou, A., & Thiare, O. (2024). A comparative study of machine learning and deep Learning methods for flood forecasting in the Far-North region, *Cameroon. Scientific African, 23*, e02053. https://www.sciencedirect.com/science/article/pii/S2468227623005069

Edmonds, M., Grattan, J., & Michnowicz, S. (2018). Volcanic gases: Silent killers. Observing the Volcano World: Volcano Crisis Communication, 65–83. https://doi.org/10.1007/11157_2015_14

Fenger, J. (1999). Urban air quality. *Atmospheric Environment, 33*(29), 4877–4900. https://www.sciencedirect.com/science/article/abs/pii/S1352231099002903

Fiandrino, C., Zhang, C., Patras, P., Banchs, A., & Widmer, J. (2020). A machine-learning-based framework for optimizing the operation of future networks. *IEEE Communications Magazine, 58*(6), 20–25. https://ieeexplore.ieee.org/abstract/document/9141210

Fund, S. (2015). Sustainable development goals. https://www.un.org/sustainabledevelopment/inequality

Goh, C. C., Kamarudin, L. M., Zakaria, A., Nishizaki, H., Ramli, N., Mao, X., Zakaria, S. M. M. S., Kanagaraj, E., Sukor, A. S. A., & Elham, M. F. (2021). Real-time in-vehicle air quality monitoring system using machine learning prediction algorithm. *Sensors, 21*(15), 4956. https://doi.org/10.3390/s21154956

Gupta, N. S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Arulkumaran, G. (2023). Prediction of air quality index using machine learning techniques: A comparative analysis. *Journal of Environmental and Public Health, 2023*, 1–26. https://www.hindawi.com/journals/jeph/2023/4916267/

Halsana, S. (2020). Air quality prediction model using supervised machine learning algorithms. International Journal of Scientific Research in Computer Science, *Engineering and Information Technology, 8,* 190–201. https://www.academia.edu/download/76904233/CSEIT206435.pdf

Harishkumar, K. S., Yogesh, K. M., & Gad, I. (2020). Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models. *Procedia Computer Science, 171*, 2057–2066. https://www.sciencedirect.com/science/article/pii/S1877050920312060

Horton, R. E. (1940). An approach toward a physical interpretation of infiltration capacity. In *Soil science Society of America proceedings* (Vol. 5, No. 399-417, p. 24). https://doi.org/10.2136/sssaj1941.036159950005000C0075x

Hu, J., Ying, Q., Wang, Y., & Zhang, H. (2015). Characterizing multi-pollutant air pollution in China: Comparison of three air quality indices. *Environment international, 84*, 17–25. https://www.sciencedirect.com/science/article/abs/pii/S0160412015300052

Idrees, Z., Zou, Z., & Zheng, L. (2018). Edge computing based IoT architecture for low cost air pollution monitoring systems: A comprehensive system analysis, design considerations & development. *Sensors, 18*(9), 3021. https://www.mdpi.com/1424-8220/18/9/3021

Jacob, M. T., Michaux, K. N., Bertrand, B., Yvette Flore, T. S., Nasser, N., Vitrice Ruben, F. S., ... & Saïdou. (2021). Low-cost air quality monitoring system design and comparative analysis with a conventional method. *International Journal of Energy and Environmental Engineering, 12*(4), 873-884. https://link.springer.com/article/10.1007/s40095-021-00415-y

Jourdain, L., Roberts, T. J., Pirre, M., & Josse, B. (2016). Modeling the reactive halogen plume from Ambrym and its impact

on the troposphere with the CCATT-BRAMS mesoscale model. *Atmospheric Chemistry and Physics, 16*(18), 12099–12125. https://acp.copernicus.org/articles/16/12099/2016/

Juarez, E. K., & Petersen, M. R. (2021). A comparison of machine learning methods to forecast tropospheric ozone levels in Delhi. *Atmosphere, 13*(1), 46. https://www.mdpi.com/2073-4433/13/1/46

Kalaivani, G., & Mayilvahanan, P. (2021). Air quality prediction and monitoring using machine learning algorithm based IoT sensor-a researcher's perspective. In 2021 6th International Conference on Communication and Electronics Systems (ICCES) (pp. 1–9). IEEE. https://ieeexplore.ieee.org/abstract/document/9489153

Kelly, F. J., Fuller, G. W., Walton, H. A., & Fussell, J. C. (2012). Monitoring air pollution: Use of early warning systems for public health. *Respirology, 17*(1), 7-19. https://onlinelibrary.wiley.com/doi/full/10.1111/j.1440-1843.2011.02065.x

Kramer, O., & Kramer, O. (2013). K-nearest neighbors. Dimensionality reduction with unsupervised nearest neighbors, 13–23. https://doi.org/10.1007/978-3-642-38652-7_2

Kumari, S., & Jain, M. K. (2018). A critical review on air quality index. Environmental Pollution: Select Proceedings of ICWEES-2016, 87–102. https://doi.org/10.1007/978-981-10-5792-2_8

Kumari, N. A., Kumar, K. A., Raju, S. H. V., Vasuki, H. R., & Nikesh, M. P. (2020). Prediction of air quality in industrial area. In 2020 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT) (pp. 193–198). IEEE. https://ieeexplore.ieee.org/abstract/document/9315660

Lei, M. T., Monjardino, J., Mendes, L., Gonçalves, D., & Ferreira, F. (2019). Macao air quality forecast using statistical methods. *Air Quality, Atmosphere & Health, 12*, 1049-1057. https://link.springer.com/article/10.1007/s11869-019-00721-9

Lei, M. T., Monjardino, J., Mendes, L., Gonçalves, D., & Ferreira, F. (2020). Statistical forecast of pollution episodes in Macao during national holiday and COVID-19. *International Journal of Environmental Research and Public Health, 17*(14), 5124. https://www.mdpi.com/1660-4601/17/14/5124

Lei, T. M., Siu, S. W., Monjardino, J., Mendes, L., & Ferreira, F. (2022). Using machine learning methods to forecast air quality: A case study in Macao. *Atmosphere, 13*(9), 1412. https://www.mdpi.com/2073-4433/13/9/1412

Leong, W. C., Kelani, R. O., & Ahmad, Z. (2020). Prediction of air pollution index (API) using support vector machine (SVM). *Journal of Environmental Chemical Engineering, 8*(3), 103208. https://www.sciencedirect.com/science/article/abs/pii/S2213343719303318

Li, L., Li, Q., Huang, L., Wang, Q., Zhu, A., Xu, J., ... & Chan, A. (2020). Air quality changes during the COVID-19 lockdown over the Yangtze River Delta Region: An insight into the impact of human activity pattern changes on air pollution variation. *Science of the Total Environment, 732*, 139282. https://www.sciencedirect.com/science/article/pii/S0048969720327996

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: A review. *Frontiers in Public Health, 8*, 505570. https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2020.00014/full

Martín-Baos, J. Á., Rodriguez-Benitez, L., García-Ródenas, R., & Liu, J. (2022). IoT based monitoring of air quality and traffic using regression analysis. *Applied Soft Computing, 115*, 108282. https://www.sciencedirect.com/science/article/abs/pii/S1568494621010917

Martínez, N. M., Montes, L. M., Mura, I., & Franco, J. F. (2018). Machine learning techniques for PM 10 levels forecast in Bogotá. In 2018 ICAI Workshops (ICAIW) (pp. 1–7). IEEE. https://ieeexplore.ieee.org/abstract/document/8554995

Mishra, D., & Goyal, P. (2015). Analysis of ambient air quality using fuzzy air quality index: A case study of Delhi, India. *International Journal of Environment and Pollution, 58*(3), 149-159. https://www.inderscienceonline.com/doi/abs/10.1504/IJEP.2015.077173

Molina, M. J., & Molina, L. T. (2004). Megacities and atmospheric pollution. *Journal of the Air & Waste Management Association, 54*(6), 644-680. https://www.tandfonline.com/doi/abs/10.1080/10473289.2004.10470936

Momo, M. N., Beauvais, A., Tematio, P., & Yemefack, M. (2020). Differentiated Neogene bauxitization of volcanic rocks (western Cameroon): Morpho-geological constraints on chemical erosion. *Catena, 194*, 104685. https://www.sciencedirect.com/science/article/abs/pii/S0341816220302356

Morapedi, T. D., & Obagbuwa, I. C. (2023). Air pollution particulate matter (PM2. 5) prediction in South African cities using machine learning techniques. *Frontiers in Artificial Intelligence, 6*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10595005/

Nducol, N., Siaka, Y. F. T., Yakum-Ntaw, S. Y., Saidou, Manga, J. D., & Vardamides, J. C. (2021a). Preliminary study of black carbon content in airborne particulate matters from an open site in the city of Yaoundé, Cameroon. *Environmental Monitoring and Assessment, 193*, 1-11. https://doi.org/10.1007/s10661-021-08924-3

Nducol, N., Tchuente Siaka, Y. F., Younui Yakum-Ntaw, S., Saidou, Dika Manga, J., Vardamides, J. C., ... & Simo, A. (2021b). Ambient air particle mass concentrations in the urban area of the capital city of Yaoundé (Cameroon, Central Africa): Monthly and seasonal behaviour. *International Journal of Environmental Analytical Chemistry, 101*(15), 2909-2925. https://doi.org/10.1080/03067319.2020.1715378

Ni, K., Ramanathan, N., Chehade, M. N. H., Balzano, L., Nair, S., Zahedi, S., Kohler E., Pottie G., Hansen M., Srivastava, M. (2009). Sensor network data fault types. *ACM Transactions on Sensor Networks (TOSN), 5*(3), 1-29. https://dl.acm.org/doi/abs/10.1145/1525856.1525863

Omer, A. M. (2008). Energy, environment and sustainable development. *Renewable and Sustainable Energy Reviews, 12*(9), 2265–2300. https://www.sciencedirect.com/science/article/abs/pii/S1364032107000834

Pradeep, A. K., Appel, A., & Sthanunathan, S. (2018). AI for marketing and product innovation: Powerful new tools for predicting trends, connecting with customers, and closing sales. *John Wiley & Sons*. https://worldcat.org/title/1027563215

Pucher, J., Peng, Z. R., Mittal, N., Zhu, Y., & Korattyswaroopam, N. (2007). Urban transport trends and policies

in China and India: Impacts of rapid economic growth. *Transport Reviews, 27*(4), 379-410. https://www.tandfonline.com/doi/abs/10.1080/01441640601089988

Ray, S., & Ray, I. A. (2011). Impact of population growth on environmental degradation: Case of India. *Journal of Economics and Sustainable Development, 2*(8), 72–77. https://www.iiste.org/

Rybarczyk, Y., & Zalakeviciute, R. (2018). Machine learning approaches for outdoor air quality modelling: A systematic review. *Applied Sciences, 8*(12), 2570. https://www.mdpi.com/2076-3417/8/12/2570

Sai, K. B. K., Subbareddy, S. R., & Luhach, A. K. (2019). IOT based air quality monitoring system using MQ135 and MQ7 with machine learning analysis. *Scalable Computing: Practice and Experience, 20*(4), 599–606. https://www.scpe.org/index.php/scpe/article/view/1561

Sensortech, S. (2015). The MiCS-6814 is a compact MOS sensor with three fully independent sensing elements on one package. https://sensorsandpower.angst-pfister.com/

Sharma, M., Jain, S., Mittal, S., & Sheikh, T. H. (2021). Forecasting and prediction of air pollutants concentrates using machine learning techniques: The case of India. *In IOP Conference Series: Materials Science and Engineering, 1022*(1), 012123. IOP Publishing. https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012123/meta

Sigurdsson, H. (1988). Gas bursts from Cameroon crater lakes: A new natural hazard. *Disasters, 12*(2), 131-146. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-7717.1988.tb00661.x

Sonawani, S., & Patil, K. (2024). Air quality measurement, prediction and warning using transfer learning based IOT system for ambient assisted living. *International Journal of Pervasive Computing and Communications, 20*(1), 38-55. https://www.emerald.com/insight/content/doi/10.1108/IJPCC-07-2022-0271/full/html

Soni, H. B., & Patel, J. (2017). Assessment of ambient air quality and air quality index in golden corridor of Gujarat, India: A case study of Dahej Port. *International Journal of Environment, 6*(4), 28–41.

Su, Y. (2020). Prediction of air quality based on gradient boosting machine method. In 2020 International Conference on Big Data and Informatization Education (ICBDIE) (pp. 395–397). IEEE. https://ieeexplore.ieee.org/abstract/document/9150155

Sun, A. Y., & Scanlon, B. R. (2019). How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. *Environmental Research Letters, 14*(7), 073001. https://iopscience.iop.org/article/10.1088/1748-9326/ab1b7d/meta

Usmani, M. (2022). *Development of remote sensing-based predictive analytics to understand complex environmental problems for ensuring sustainability of human well-being (Doctoral dissertation, University of Florida)*. https://original-ufdc.uflib.ufl.edu/UFE0058470/00001

Venkatraman Jagatha, J., Klausnitzer, A., Chacón-Mateos, M., Laquai, B., Nieuwkoop, E., van der Mark, P., & Schneider, C. (2021). Calibration method for particulate matter low-cost sensors used in ambient air quality monitoring and research. *Sensors, 21*(12), 3960. https://www.mdpi.com/1424-8220/21/12/3960

World Health Organization. (2010). *WHO guidelines for indoor air quality: Selected pollutants. World Health Organization*. Regional Office for Europe. https://iris.who.int/bitstream/handle/10665/260127/9789289002134-eng.pdf

Yaulande, D. A., André, D., Ossénatou, M., & André, L. (2022). *Prediction of daily direct solar energy based on xgboost in Cameroon and key parameter impacts analysis*. In 2022 IEEE Multi-conference on Natural and Engineering Sciences for Sahel's Sustainable Development (MNE3SD) (pp. 1–7). IEEE. https://ieeexplore.ieee.org/abstract/document/9723309

Yu, R., Yang, Y., Yang, L., Han, G., & Move, O. A. (2016). RAQ–A random forest approach for predicting air quality in urban sensing systems. *Sensors, 16*(1), 86. https://www.mdpi.com/1424-8220/16/1/86

Yuh, Y. G., Tracz, W., Matthews, H. D., & Turner, S. E. (2023). Application of machine learning approaches for land cover monitoring in northern Cameroon. *Ecological informatics, 74*, 101955. https://www.sciencedirect.com/science/article/pii/S1574954122004058

Zhang, D., Du, L., Wang, W., Zhu, Q., Bi, J., Scovronick, N., ... & Liu, Y. (2021). A machine learning model to estimate ambient PM2. 5 concentrations in industrialized highveld region of South Africa. *Remote sensing of environment, 266*, 112713. https://www.sciencedirect.com/science/article/abs/pii/S0034425721004338

Zhu, S., Lian, X., Liu, H., Hu, J., Wang, Y., & Che, J. (2017). Daily air quality index forecasting with hybrid models: A case in China. *Environmental Pollution, 231*, 1232–1244. https://www.sciencedirect.com/science/article/abs/pii/S0269749117316330