MethodsX 14 (2025) 103322

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/methodsx

Methodological Integration of Machine Learning and Geospatial Analysis for PM_{10} Pollution Mapping

Kalid Hassen Yasin[a,*], Muaz Ismael Yasin[b], Anteneh Derribew Iguala[c], Tadele Bedo Gelete[a], Erana Kebede[d]

[a] *Geo-Information Science Program, School of Geography and Environmental Studies, Haramaya University, P.O. Box 138, 3220 Dire Dawa, Ethiopia*
[b] *School of Medicine, College of Health and Medical Sciences, Haramaya University, P.O. Box 235, Harar, Ethiopia*
[c] *Center for Rural Development, Oromia State University, Batu P.O. Box 209, Ethiopia*
[d] *School of Plant Sciences, College of Agriculture and Environmental Sciences, Haramaya University, P.O. Box 138, Dire Dawa, Ethiopia*

A R T I C L E I N F O

Method name:
NB, RF, and KNN

Keywords:
Air Pollution
Pollution
Data-driven approach
Spatial prediction

A B S T R A C T

Air pollution mitigation necessitates accurate spatial modelling to inform public health interventions. Traditional approaches inadequately capture complex predictor-pollutant interactions, whereas machine learning (ML) offers a superior capacity for modelling nonlinear relationships. This study compares three ML Random Forest (RF), K-Nearest Neighbors (KNN), and Naïve Bayes (NB) algorithms using annual PM_{10} data from 11 monitoring stations alongside atmospheric, urban, and terrain covariates. The methodological framework employed rigorous preprocessing and cross-validation to classify pollution into three categorical levels. Results demonstrate RF superior performance, achieving 94% balanced accuracy and 97% specificity, significantly outperforming KNN (92%) and NB (89%). RF excelled in capturing spatial heterogeneity and complex variable interactions, while KNN and NB exhibited limitations in managing feature dependencies and localized variability. Despite computational demands, findings substantiate RF reliability for robust air quality monitoring applications. The study contributes valuable insights for implementing scalable pollution prediction systems in resource-constrained urban environments while acknowledging interpretability challenges inherent to complex ML models.

- Preprocessing of spatial data from various sources, incorporating the handling of missing/abnormal data, analysis, and normalization
- Implementation of the three ML algorithms with rigorous hyperparameter tuning, model validation, and performance assessment
- Mapping PM_{10} Hotspots on the Gradient Direction and Distance from the City Center

* Corresponding author.
E-mail addresses: kalid.hassen@haramaya.edu.et, kalidh84@gmail.com (K.H. Yasin).

https://doi.org/10.1016/j.mex.2025.103322
Received 10 January 2025; Accepted 16 April 2025
Available online 17 April 2025

Specifications table

Subject area:	Environmental Science, Earth and Planetary Sciences and Computer Science
More specific subject area:	Machine Learning, Air Pollution
Name of your method:	NB, RF, and KNN
Name and reference of the original method:	K. H. Yasin, M. I. Yasin, A. D. Iguala, T. B. Gelete, D. Tulu, E. Kebede, Predictive machine learning and geospatial modeling reveal PM_{10} hotspots and guide targeted air pollution interventions in Addis Ababa, Ethiopia, Discover Applied Sciences 7 (4) (2025) 263. https://doi.org/10.1007/s42452-025-06723-w.
Resource availability:	Repository name: Zenodo Data identification number: 10.5281/zenodo.12825035 Direct URL to data: https://zenodo.org/records/12825036 Code availability: https://github.com/Dodokal/PM10-Prediction

Background

Particulate matter PM_{10} (<10 μm) constitutes a significant environmental hazard with established pathophysiological sequelae, penetrating respiratory passages to induce inflammatory cascades associated with pulmonary dysfunction, cardiovascular pathology, and elevated mortality indices [1]. Precise mapping and spatial prediction of PM_{10} concentrations are essential for environmental monitoring and the application of successful mitigation techniques [2]. Traditional approaches to modelling air quality, such as dispersion and land use regression models, are limited in their ability to capture complex interactions between different predictor variables and their nonlinear relationship to PM_{10} [3]. Moreover, these methods often rely on assumptions and simplifications that may not accurately reflect the reality of the situation in the real world.

Machine learning (ML) algorithms provide potent alternatives for spatial prediction of PM_{10} concentrations [2]. These data-driven techniques can learn from large data sets, identify complex patterns, and model nonlinear relationships without relying on underlying processes to be robust [4]. The proposed methodology leverages three widely used machine learning algorithms (MLAs), Naïve Bayes (NB), Random Forest (RF), and K-Nearest Neighbor (KNN), to study their performance in predicting PM_{10} pollution levels (categorized as Good, Moderate, and Unhealthy for Sensitive Groups (UnHealSens)) in Addis Ababa, Ethiopia. By comparing the performance of these different ML algorithms, the study aims to determine the most appropriate approach to spatial prediction of PM_{10} levels. The comprehensive evaluation framework, including data preprocessing, hyperparameter tuning, and rigorous model validation, ensures the robustness and generalizability of the results. The novelty and contribution to this field is that this is the first study in Addis Ababa, Ethiopia, to predict PM_{10} levels and map hotspot areas with a comparative MLA analysis.

Method details

Data record preprocessing and data entry

The study used PM_{10} concentration data obtained from the World Air Quality Index project database (https://waqi.info/ or https://aqicn.org/) (Fig. 1). Due to the limited data availability, the study used the annual average PM_{10} values ($\mu g/m^3$) for 11 measuring stations for the period August 2021 – August 2023. Using annual averages, we aimed to reduce short-term fluctuations and focus on long-term trends in pollution, thus ensuring a stable input for the local and metropolitan authorities. The initial data collection was done in monthly Excel spreadsheets, followed by averaging for practical data analysis. The collected data points and

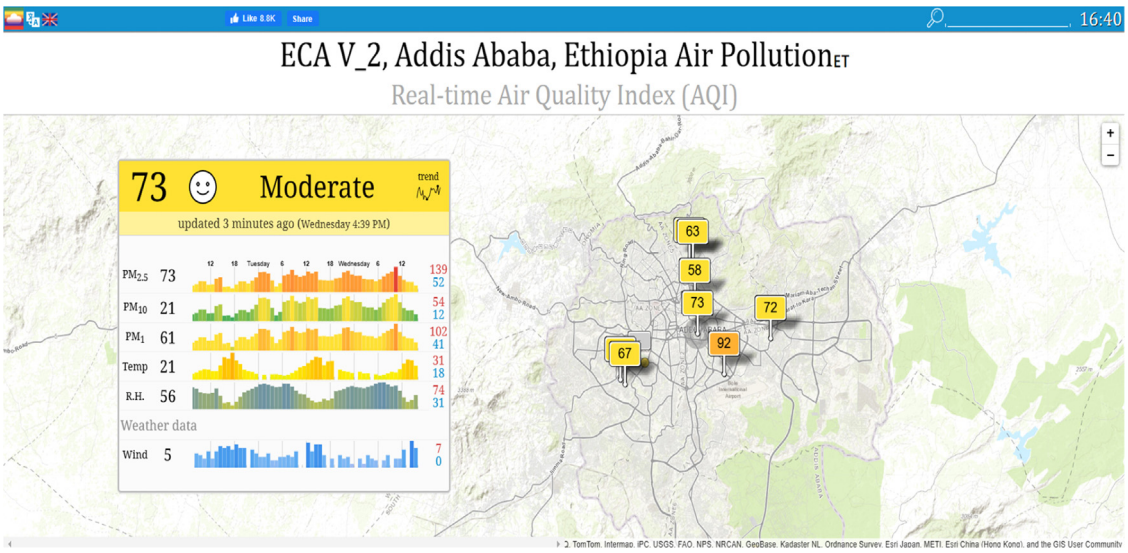

Fig. 1. Screenshot of the WAQI database for PM_{10} monitoring station interface in Addis Ababa.

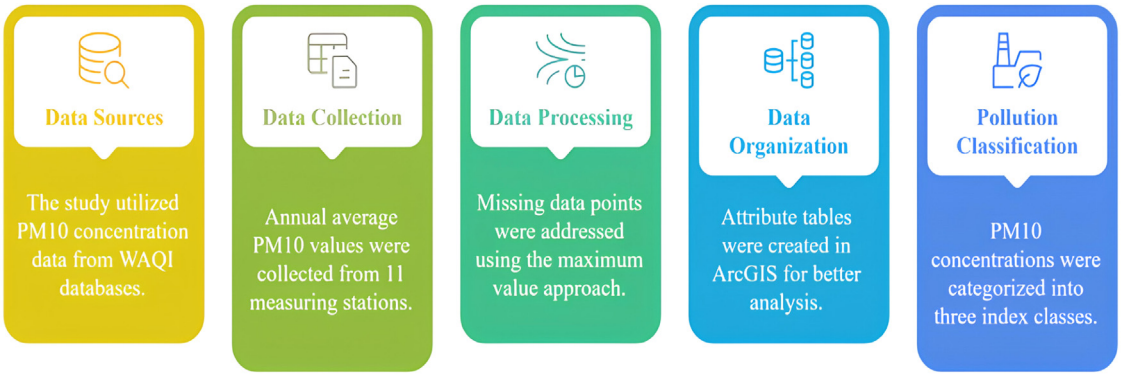

Fig. 2. PM_{10} concentration analysis workflow, from data sourcing to pollution classification.

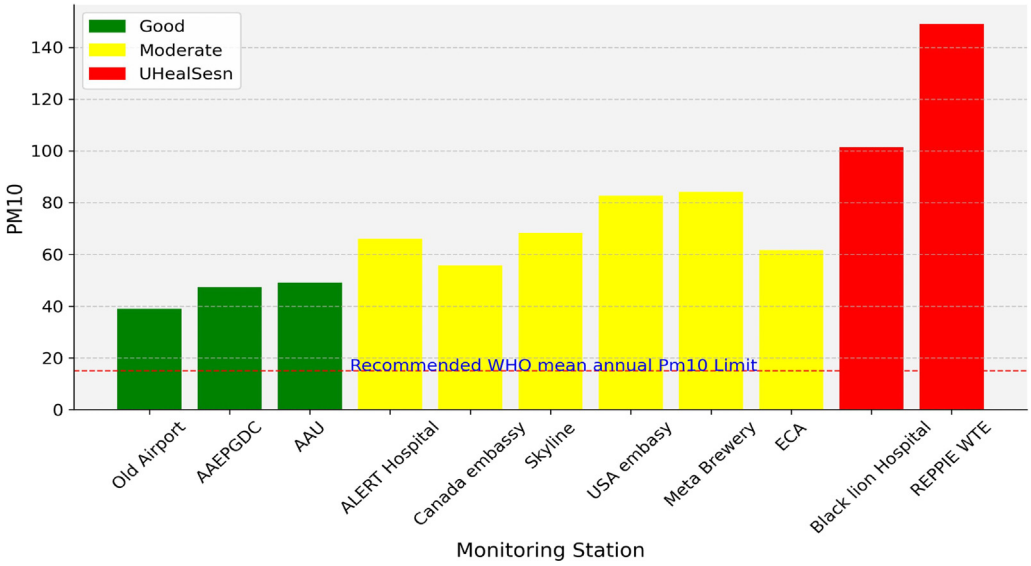

Fig. 3. Mean PM_{10} distribution by the monitoring station. The red dotted line at 15 μg/m^3 is the annual mean WHO guideline.

the corresponding station locations have been imported into ArcGIS for further spatial analysis. The main problem with the use of high temporal resolution data (daily or monthly PM_{10} values) was the existence of missing data from several stations, making it difficult to assess the time differences accurately. We used a maximum value approach to address this, replacing missing data points with the highest observed concentration over a comparable period. This approach was chosen to ensure that the pollution level is not underestimated while preserving the integrity of the data. Unlike average or interpolated estimates, the maximum value approach favours conservative estimates. It helps to capture peak pollution events, which are often more relevant to the health assessment of an environment. This decision was made to preserve the data's integrity and avoid inaccuracies in the methods of averaging or estimation. A new attribute field named *'Station'* was included to identify each monitoring station uniquely (Fig. 2).

Concentration values were entered into the dataset using the actual data types to ensure accuracy, while the '*levels*' field was created to categorize the pollution levels according to string data types, allowing for categorical classification. We have entered specific concentration values for the different stations (Fig. 3). In addition, PM_{10} concentrations have been broken down into three index classes using flexible margins as defined by the World Health Organization (WHO) [2].

Validation Framework for PM_{10} Sensor Data

The reliability of the sensor data used for PM_{10} prediction has been assured by implementing a rigorous data validation process (Fig. 4). This methodological approach included separate quality assessment and comparative validation with co-located sensors, following the WAQI standards. The Signal-to-Noise Ratio (SNR) served as the primary metric for data reliability assessment, quantifying the variability of sensor data to the mean values. Well-maintained sensors show relatively low SNR values, while excessive noise in the data may indicate measurement errors due to sensor faults or environmental interference [5].

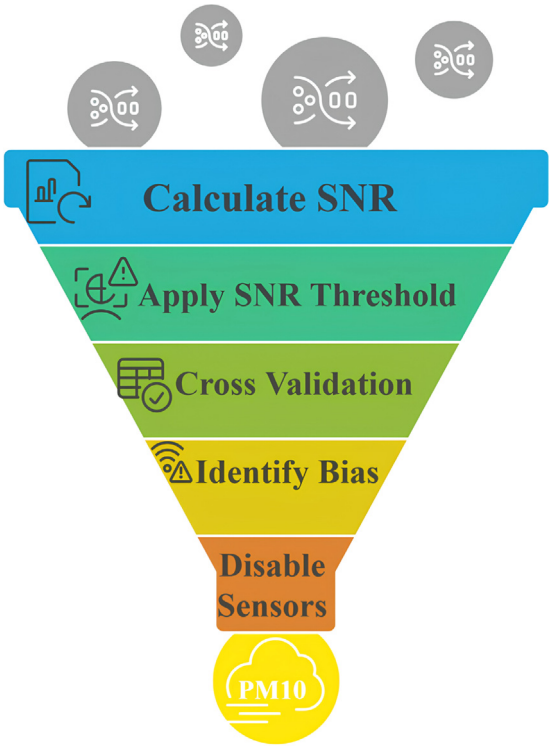

Fig. 4. WAQI's SNR-based sensor filtering process for PM_{10} measurement involves SNR calculation, threshold application, cross-validation, bias identification, and sensor disabling.

The calculation of SNR employed the coefficient of variation (CV), which is mathematically represented as Eq. 1:

$$SNR = \frac{\sigma}{\mu} \times 100\% \tag{1}$$

where σ is the standard deviation of the hourly PM_{10} readings, which measures the degree of fluctuation in recorded values over a given period. μ denotes the mean PM_{10} concentration over the same period.

In order to maintain high data quality standards, WAQI has set a threshold of 33 % SNR as the upper acceptable limit, beyond which data are considered unreliable and are therefore excluded from further analysis. In addition, all sensors with a consistent SNR of more than 10 % were systematically flagged for possible bias, as these patterns may indicate mechanical failures, such as a faulty fan, or external factors, including dust accumulation in the sensor housing. In addition to the separate quality assessment based on the SNR, the validation protocol included a comparative validation approach to confirm the consistency of sensor data with the co-located sensors in the monitoring network. This supplementary method is based essentially on a confidence interval of the data readings, defined explicitly as three times the standard deviation (3σ) of the hourly PM_{10} values from adjacent stations. The underlying assumption is that valid sensor data should correspond closely to the average trend of nearby stations unless there is a valid local pollution event.

A station was considered to generate abnormal readings if its hourly PM_{10} concentration consistently exceeded the confidence zone. The probability of a sensor producing inaccurate data was quantified using the following weight function (Eq. 2):

$$d_i = \frac{v_i - median_i}{stddev_i},\ P = \sum_i \left(\frac{1}{n},\ d_i > 5; 0,\ otherwise\right),\ D = \sqrt{\frac{\sum_i d_i}{n}} \tag{2}$$

where d_i represents the normalized deviation, calculated as the difference between the sensor's hourly PM_{10} reading (v_i) and the median of neighboring stations ($median_i$), divided by the standard deviation of those neighboring stations ($stddev_i$). P represents the probability of deviation beyond the confidence zone, where each instance of $d_i > 5$ contributes a value of $\frac{1}{n}$, and all other instances contribute zero. D is the statistical distance of the station's reading from the median of its neighbors, calculated as the square root of the mean deviation over n samples. n is the number of hourly readings over the past three days. This function identifies instances where a station consistently deviates significantly from nearby stations, using a threshold of $d_i > 5$.

The final weight function was computed using Eq. 3:

$$W = P \times D \tag{3}$$

If the weight function W exceeds 30, the station is considered unreliable and is automatically shut down to avoid skewing the overall data of PM_{10}.

Table 1
Source of data collection for PM_{10} Predictor variables.

Predictor Variables	Resolution or scale	Source
PM_{10} Concentration	CSV	WAQI (https://waqi.info or https://aqicn.org)
AD, WEX, WS	250 × 250 m; resampled to 30 × 30 m	Global Wind Atlas (https://globalwindatlas.info/)
Ta, RF	0.93 × 0.93 km; resampled to 30 × 30 m	Worldclim (https://www.worldclim.org/)
NL	15 arc second; resampled to 30 × 30 m	VIIRS (https://eogdata.mines.edu/)
LST, SAVI, UI,	10 × 10 m; resampled to 30 × 30 m	SENTINEL-2A (https://scihub.copernicus.eu/)
ELV, SLP, TWI, TRI,	30 × 30 m	Copernicus DSM (https://spacedata.copernicus.eu/)
BD	Shapefiles (1:25000)	Open Buildings (https://sites.research.google/open-buildings)
Road, Airport, Railway, Construction sites	Shapefiles (1:25000)	OpenStreetMap (OSM) (https://www.openstreetmap.org)
Petrol stations, Quarries, Unions and marketplaces, Wasteplate and Landfill, Bus-station, industries	Shapefiles (1:25000)	SSGI Geo-Portal (http://www.ethionsdi.gov.et/)

The validation framework is accessible through the WAQI validation dashboard (https://aqicn.org/data-feed/validation/), which visually represents historical and real-time SNR assessments. Additionally, researchers can verify sensor reliability using station-specific identifiers through the WAQI data-feed validation platform, which facilitates validation based on unique station names or identification numbers (e.g., https://aqicn.org/station/validation/@204991). These automated validation processes ensure better data quality and minimize sensor bias in PM_{10} measurements.

Covariate processing

Covariate processing for the PM_{10} analysis adopted a comprehensive approach that included four main categories of predictor variables adopted from previous studies [2,6,7]: atmospheric conditions, urban characteristics, terrain characteristics, and proximity measurements to potential PM_{10} sources. This diverse dataset provides essential insights into the environmental and urban factors influencing PM_{10} concentrations and enables a nuanced understanding of air pollution dynamics. The predictor variables were obtained from various open data sources and organizations, as listed in Table 1.

Atmospheric conditions, a key part of the analysis, included variables such as atmospheric density (AD), temperature (Ta), the wind exposure index (WEX), wind speed (WS), and precipitation (RF). These parameters, sourced from the Global Wind Atlas and Worldclim, initially had different resolutions but were converted to a unified 30 × 30 m grid to ensure consistency across the dataset. This standardization allowed for a more accurate representation of the atmospheric processes that impact PM_{10} levels. Urban characteristics, another important category, included night light (NL) data, land surface temperature (LST), urban indices (UI), road density, and building density (BD). These urban activity and infrastructure indicators were derived from various sources, including the VIIRS dataset, Sentinel-2A images, and Open Buildings shapefiles. The integration of these urban parameters enabled a comprehensive assessment of the built environment's impact on the PM_{10} concentration.

The terrain characteristics, which included elevation (ELV), slope (SLP), topographic wetness index (TWI), and terrain roughness index (TRI), were processed using Copernicus DSM data. These variables, maintained at a resolution of 30 × 30 m, provided crucial information about how topographic features influence the spread and concentration of PM_{10} and deepened the analysis of pollution patterns. Proximity measurements of potential PM_{10} sources were carefully collected from various repositories, including OpenStreetMap and the Space Science and Geospatial Institute Geo-Portal. These measurements, which included distances to airports, petrol stations, roads, and other potential sources of pollution, were crucial in assessing local impacts on PM_{10} levels. The data processing workflow included multiple phases and used the Google Earth Engine (GEE) and ArcGIS. In the GEE, indices such as the SAVI, UI, and LST were processed to ensure temporal and spatial alignment. ArcGIS was used to process climatic variables, proximity measurements, and terrain characteristics, which requires complex spatial analysis and buffering techniques. To ensure consistency and allow for accurate modelling, all covariate raster images were masked to the size of the study area and projected onto the WGS84 UTM 37 coordinate system, then thoroughly cleaned to remove any redundant information and discrepancies. This rigorous methodology has resulted in a reliable, accurate, and well-structured data set for modelling PM_{10} level and has provided a solid basis for analyzing environmental pollution. The relationship between PM_{10} levels and the above-mentioned predictor variables is illustrated in Fig. 5, which gives a clear picture of the complex interactions.

Training Data Preparation

Data Extraction

Preparing the training data for air pollution analysis requires careful data extraction, cleaning, and normalization. PM_{10} concentration data were obtained from 11 monitoring stations across the city, as these were the only available stations that met data validity and quality assurance criteria. Given the limited number of monitoring locations, a data augmentation approach was applied to enhance the spatial representativeness of the dataset. This expansion required a strategic decision between reducing the size of the study area for more detailed data collection or increasing the number of sampling sites, depending on the available resources. The entire area was used, considering the potential uncertainties associated with the data augmentation method. Circular buffers with a radius of 2 km were created around each air quality monitoring station to estimate the area of influence around each monitoring station. This approach was based on the assumption that the station values could be reliably applied within this radius; beyond

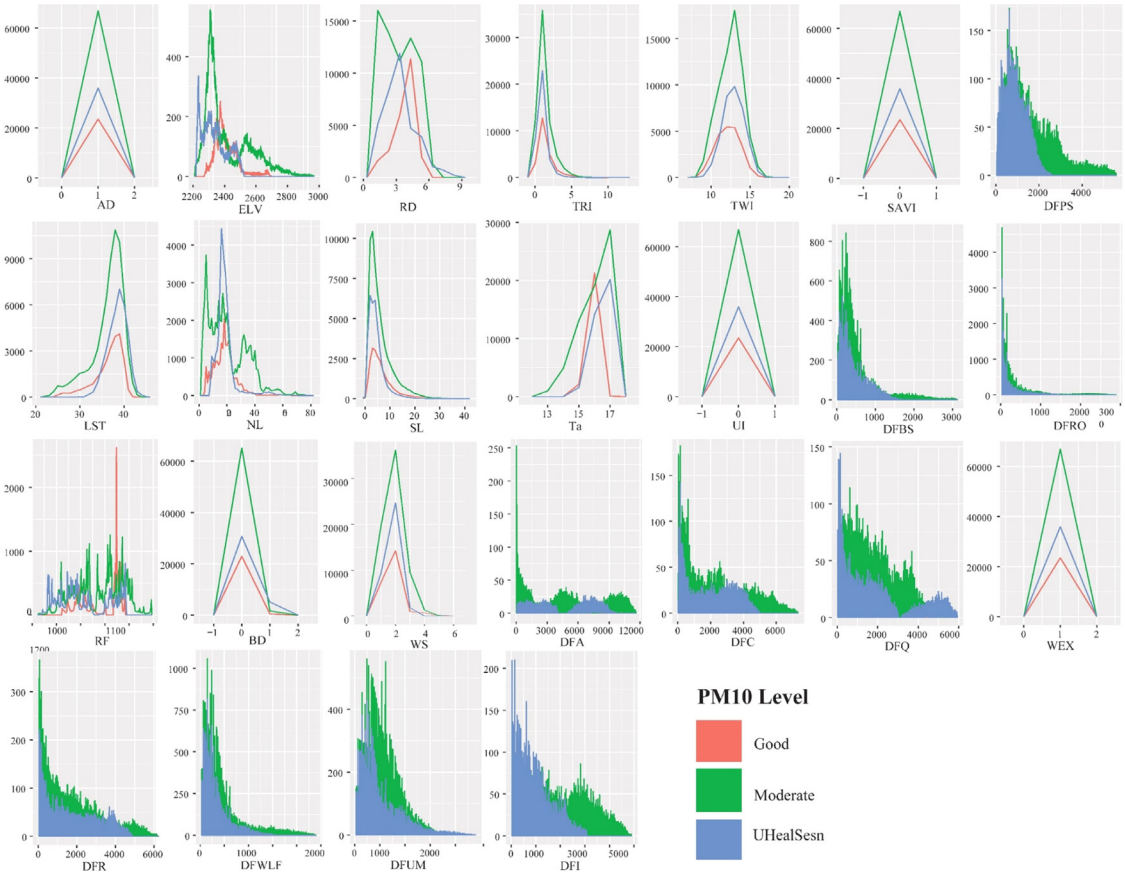

Fig. 5. Descriptive analysis showing the relationship between modelling variables and PM_{10} values (X-axis shows values, and Y-axis shows count).

this distance, wind patterns could influence the distribution of the particles and make them inhomogeneous. Using buffers with a length of 2 km was justified because larger buffers may not accurately capture local variations [2]. A medium grid cell resolution of 30 × 30 m was then adopted to balance model precision and computational power across the study area, which was aligned with the created buffers and tailored to focus on relevant regions (Fig. 6).

PM_{10} data and covariates were extracted from raster data, with pollution levels reclassified into good, moderate, and UnHealSens. QGIS facilitated joining attribute tables from the clipped data, ensuring accurate values for each grid cell. Raster statistics, particularly mean values, were calculated to gather information on conditioning factors, preparing the dataset for subsequent cleaning and normalization (Fig. 7). The predictor variables were retrieved via the Raster Statistics to Polygon tool in SAGAGIS and saved in CSV format, yielding 39,885 observations for subsequent analysis in the R environment.

Data Cleaning, Normalization, and Final Processing

R Statistical Software (v4.4.0; R Core Team 2024) was used to process PM_{10} data and predictor variables further. Cleaning the dataset reduced noise and eliminated unexpected observations, resulting in a refined dataset of 38,408 observations. Unnecessary columns have been removed, and the remaining issues have been addressed in a thorough final review. PM_{10} prediction data imported from CSV file have rows with missing values removed to ensure the completeness of the data. The data set was then converted into a data frame format containing only relevant variables for the analysis, such as atmospheric conditions, urban characteristics, terrain features, proximity measurements and the target PM_{10} level. The data normalization, which is critical for ML estimation, was achieved by a user-defined function that scaled each predictor variable to a range of 0 – 1, except for the PM_{10}-level column, which was managed as a categorical variable. Normalized data was then combined with the '*LevelAve*' column to preserve the structure of the data file. This approach resulted in a comprehensive, normalized and well-formatted data set, which was stored in a new CSV file ready for further analysis and development of a model for predicting PM_{10} levels. The final dataset was divided into a 30% testing subset with 11,522 observations and a 70% training subset with 26,886 observations. This partitioning approach was implemented to minimize overfitting and underfitting [8].

The data processing code is publicly available on Zenodo. The datasets in the repository are organized into a structured folder [9]. The folder contains four key files, each performing a specific function in the data analysis. The "*Analyzed Data*" file is one of these files and represents the unfiltered data set in CSV format, which includes 39,885 observations. This raw data provides a complete

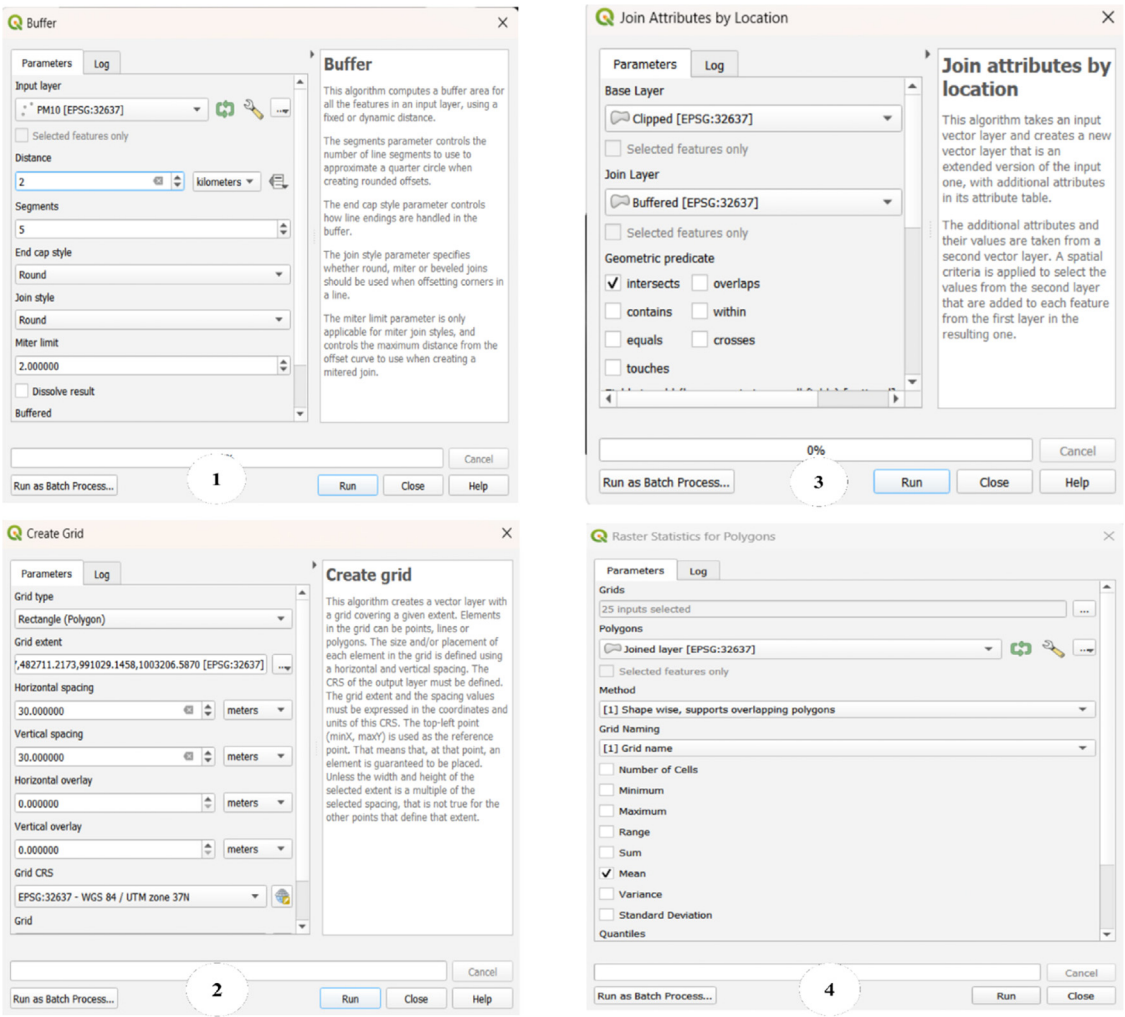

Fig. 6. Data Extraction Process.

overview of the information collected before any processing or purification steps have been carried out. Following the raw data, the "*Filtered Normalized Data*" file contains the cleaned and normalized data set. This file represents a crucial step in the data preparation process as it has been curated to eliminate redundancies and inconsistencies. The resulting dataset consists of 38,408 observations in CSV format. This cleaned data set is the basis for subsequent analysis and modelling efforts. The folder also contains a "CODE.R" file to ensure transparency and reproducibility. This R script contains the code used for data cleaning and normalization so that other researchers can understand and potentially replicate the data preparation process. A README.txt file has been included in conjunction with this code, which contains detailed explanations of the code structure and functionality and serves as a guide for those who wish to delve deeper into the data processing methodology.

Machine learning model development

The selection of MLAs for PM_{10} level prediction was based on the sample size, response type, and their proven efficacy in environmental modelling. Implemented using the CARET package in R, these algorithms facilitate essential tasks such as data preprocessing, sample selection, variable selection, data partitioning, and model construction. Their methodological robustness and strong theoretical foundations make them well-suited for capturing complex atmospheric patterns, as supported by extensive literature in air pollution modelling [2,10–26]. The descriptions of the models are presented below.

(a) RF: The RF algorithm represents a methodologically robust choice for air pollution modelling, as demonstrated by its extensive application in related environmental studies [27]. Vovk et al. [28] employed RF algorithms to predict PM2.5 concentrations with high accuracy across diverse geographical contexts. Similarly, Alzu'bi et al. [29] documented the algorithm's superior performance when handling the nonlinear relationships characteristic of atmospheric systems. The algorithm's inherent resistance to overfitting through ensemble methodology provides a significant advantage when modelling complex environmental phenomena with potentially noisy measurements [2,30,31].

Fig. 7. Training data preparation process, encompassing data extraction, cleaning, normalization, and final processing.

The RF model was developed utilizing the CARET package with cross-validation for robust performance evaluation. The fundamental equation governing RF prediction can be expressed as Eq. (4):

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{4}$$

Where $\hat{f}_{rf}^{B}(x)$ represents the ensemble prediction, B denotes the total number of trees, and $T_b(x)$ indicates the prediction from the b-th tree for input vector x.

Optimization involves systematically testing different *ntree* and *mtry* values to maximize predictive accuracy. After a comprehensive analysis, we select 500 *ntree* and 5 *mtry* values as optimal parameters. These hyperparameters undergo further refinement through a 10-fold cross-validation process while maintaining a consistent *ntree* of 500. This configuration effectively balances model complexity and computational efficiency without compromising predictive power, a critical consideration for environmental prediction systems (Fig. 8).

(b) KNN: The KNN algorithm provides its methodology for classification and regression tasks. Based on the principle of lazy learning, KNN avoids assumptions about data distribution and is, therefore, particularly suitable for air pollution prediction scenarios where universal predictors are missing. Previous research by Chao et al. [32] demonstrated KNN's effectiveness in predicting urban air quality parameters, highlighting its ability to capture localized pollution patterns without requiring explicit assumptions about data distribution [33,34]. Atmakuri et al. [35] also documented KNN's superior performance in scenarios with complex spatial variability in pollution concentrations.

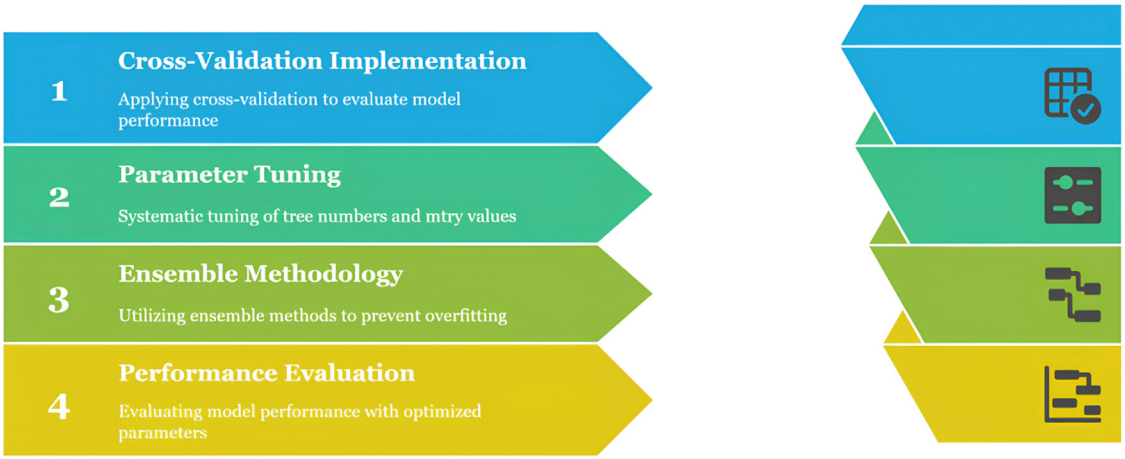

Fig. 8. Steps implemented for PM_{10} prediction using the RF model.

Fig. 9. Steps implemented for PM_{10} prediction using the KNN model.

The algorithmic workflow randomly selects k-class centers and then assigns training instances to their nearest center. Class centers undergo subsequent updates and reclassification based on the means of corresponding class data points. For new data classification, KNN identifies the k nearest neighbours and selects the class with the highest representation in the neighbourhood. The weighted classification function for KNN can be formalized as Eq. 5:

$$\hat{y} = \underset{c \in C}{\operatorname{argmax}} \sum_{i=1}^{k} w_i \cdot \mathbb{I}(y_i = c) \tag{5}$$

Where $\hat{y}$ Represents the predicted class, C denotes the set of possible classes, w_i indicates the weight assigned to the i-th neighbour, y_i is the class of the i-th neighbour, and $\mathbb{I}$ represents the indicator function that equals 1 when the argument is true and 0 otherwise.

While Euclidean distance is the primary metric for continuous features, as shown in Eq. 5, Hamming distance (Eq. 6) is more suitable for discrete features [7,36].

$$d_{euclidean} = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{6}$$

where n represents the number of dimensions, a denotes data points from the training set, and b represents new observations for prediction.

We employ k-fold cross-validation with k=10 to evaluate model performance and mitigate overfitting risks. After systematically examining various k values, we determine that k=5 yields optimal results, establishing an appropriate balance between neighbourhood size and overfitting susceptibility (Fig. 9).

(c) NB: NB classifiers based on the principles of Bayesian probability. This method assumes that the features in a class function independently, allowing each attribute to contribute consistently and independently to the probability of sampling being

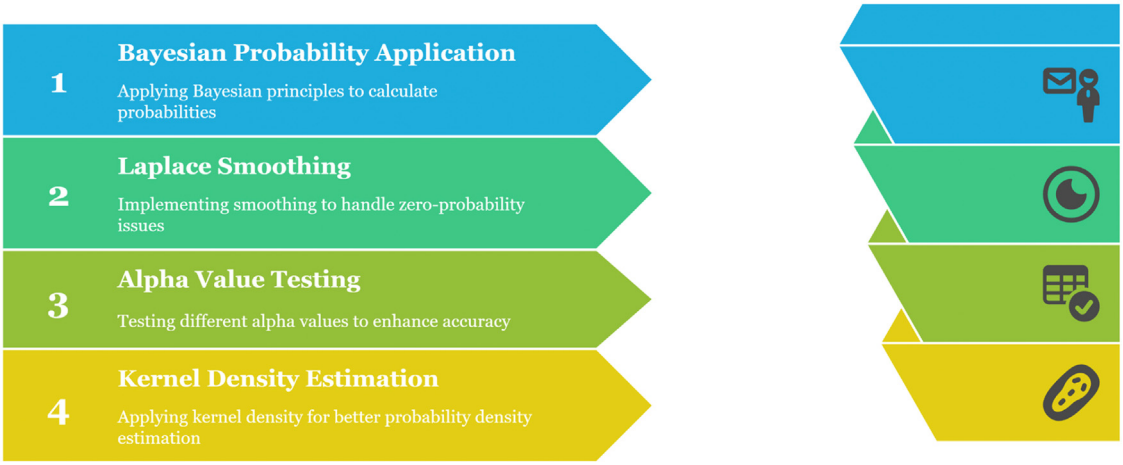

Fig. 10. Steps implemented for PM_{10} prediction using the NB model.

categorized [37,38]. NB offers compelling theoretical advantages that justify its application. A recent study by Sairam and Nagaraju [39] demonstrated NB competitive performance when predicting categorical air quality indices [22], while Pant et al. [28] highlighted the algorithm's computational efficiency when processing large environmental datasets. These characteristics render NB particularly valuable for operational air quality prediction systems requiring rapid processing. The generalized classification function for NB can be expressed as Eq. 7:

$$\hat{y} = \underset{c_k \in C}{\operatorname{argmax}} P(c_k) \prod_{i=1}^{n} P(x_i|c_k) \tag{7}$$

Where y Represents the predicted class, C denotes the set of possible classes, $P(c_k)$ indicates the prior probability of class c_k, and $P(x_i|c_k)$ represents the conditional probability of the feature x_i given class c_k.

The probabilistic foundation operates according to Eq. 8:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \tag{8}$$

where $P(c|d)$ represents the posterior probability, $P(d|c)$ the likelihood, $P(c)$ the class prior probability, and $P(d)$ the predictor prior probability.

Despite its advantages, including noise insensitivity, implementation simplicity, and indifference to dataset size, NB remains underutilized in air pollution prediction contexts [40]. Our implementation addresses potential limitations by incorporating Laplace smoothing as a correction method for zero probability values (Fig. 10). We test alpha values of 0, 0.5, and 1.0, with maximum accuracy achieved using an alpha value of 0.

The model execution utilizes the *f* function in R software, employing kernel density estimation to represent the conditional probability density function non-parametrically. We enable kernel functionality through a single bandwidth tuning parameter via the '*usekernel*' function. Significantly, bandwidth alterations substantially impact density estimation flexibility, with higher values producing denser and more flexible estimates. This parametric flexibility allows the model to adapt to the specific characteristics of the environmental data distribution, enhancing predictive performance for PM_{10} concentration levels.

The careful selection of these three algorithms establishes a methodologically sound comparative framework that addresses distinct mathematical approaches to classification. Each algorithmic implementation undergoes rigorous cross-validation and hyperparameter optimization to ensure maximum predictive accuracy while maintaining generalizability across environmental contexts. The selection of these complementary methodologies establishes a robust framework for PM_{10} prediction that leverages distinct mathematical approaches to address the inherent complexity of atmospheric PM dynamics.

Mapping PM_{10} Hotspots based on Gradient Direction and Distance

We conduct a concentric buffer analysis to identify predicted hotspots of PM_{10} pollution at various distances and in different directions from the city center, aiming to understand the spatial distribution of this pollution. The study area was systematically divided into concentric circles of 4, 8, 12, and 16.5 km in diameter; each ring was subdivided into 16 sectorial radiating outward from the city's center. This spatial segmentation has allowed for a detailed study of the gradients of pollution hotspots in radial and angular directions. The predicted PM_{10} values from the RF model, selected for their accuracy and practical applicability, were overlaid on these spatial units, allowing a detailed analysis of the pollution patterns.

The methodology consisted of several key steps. First, the ArcGIS multiple-ring buffer tool generated concentric circles centered at the city center, indicating distances of 4, 8, 12, and 16.5 km. The study area was subdivided into directionally oriented sectors

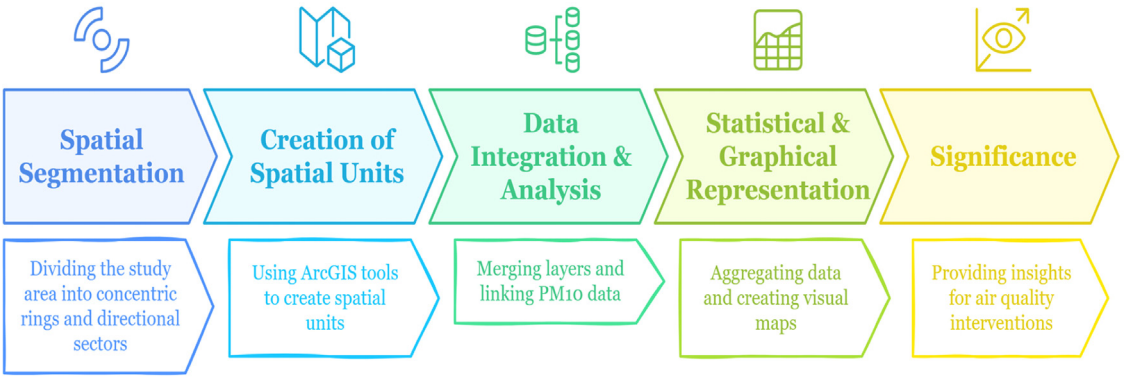

Fig. 11. Sequential Steps for Mapping PM_{10} Hotspot.

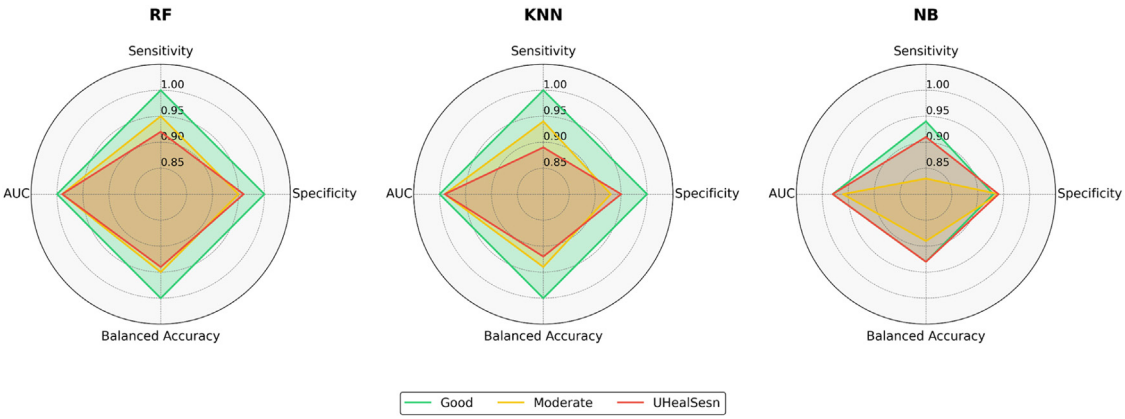

Fig. 12. Comparison of Classification Models (RF, KNN, NB) Using Radar Charts: Evaluating Sensitivity, Specificity, AUC, and Balanced Accuracy Across Different Performance Levels (Good, Moderate, and UHealSsen).

by creating a new shape file named '*Direction*' and a function type '*Polygonal*.' From a certain angle, lines representing 16 gradient directions were drawn from the city center, and these lines were converted to polygonal segments using the function '*feature to polygon tool*'. The buffer and direction sector layers have been merged using the *Union* tool with appropriate modifications to keep the required sectors and label each with direction and distance identifiers. The predicted PM_{10} emissions from the RF model have been imported into ArcGIS and linked to the generated sectors by *spatial join* or *intersection* tools. The PM_{10} levels were classified according to the pollution standards, and the field areas (in km^2) and percentages for each sector were added and calculated with the help of the *Geometry* and *Field Calculator* tools. Finally, statistical analysis was carried out using the *Summary Statistics* tool to aggregate the data on the area and percentages for each sector, and the aggregated data was exported to Excel for further analysis. Maps and graphics have been used to visualize the trends of PM_{10} in different directions and distances. This methodological approach (Fig. 11) allows a comprehensive mapping and analysis of hotspots of PM_{10} in Addis Ababa, providing key insights for targeted interventions and policy decisions to improve air quality.

Method validation

In this study, 30 % of the dataset was assigned to validation using stratified sampling to ensure representative distribution of PM_{10} concentrations. This approach preserved the proportionality of the different PM_{10} categories and minimized any bias in the model evaluation. A 10-fold cross-validation further improved the validation process to optimize the reliability and generalization of the model under different conditions. A comparison of the MLAs, RF, KNN, and NB shows various degrees of accuracy in predicting PM_{10} levels: Good, Moderate, and UnHealSens. To provide a clear and systematic assessment, the performance metrics of the classifier, including sensitivity, specificity, balanced accuracy, and area under the receiver operating characteristic curve (AUC) (Fig. 12). The RF model consistently outperformed the KNN and the NB model on all assessment parameters, confirming its robustness to handle complex environmental data sets. RF achieved an AUC of 1.00 for the Good class, which is a perfect classification, and maintained a high precision for the Moderate and UnHealSens categories (AUC = 0.99). These findings highlight the effectiveness of RF in capturing spatial variation in PM_{10} levels, which is a key requirement for monitoring and developing urban air quality policies.

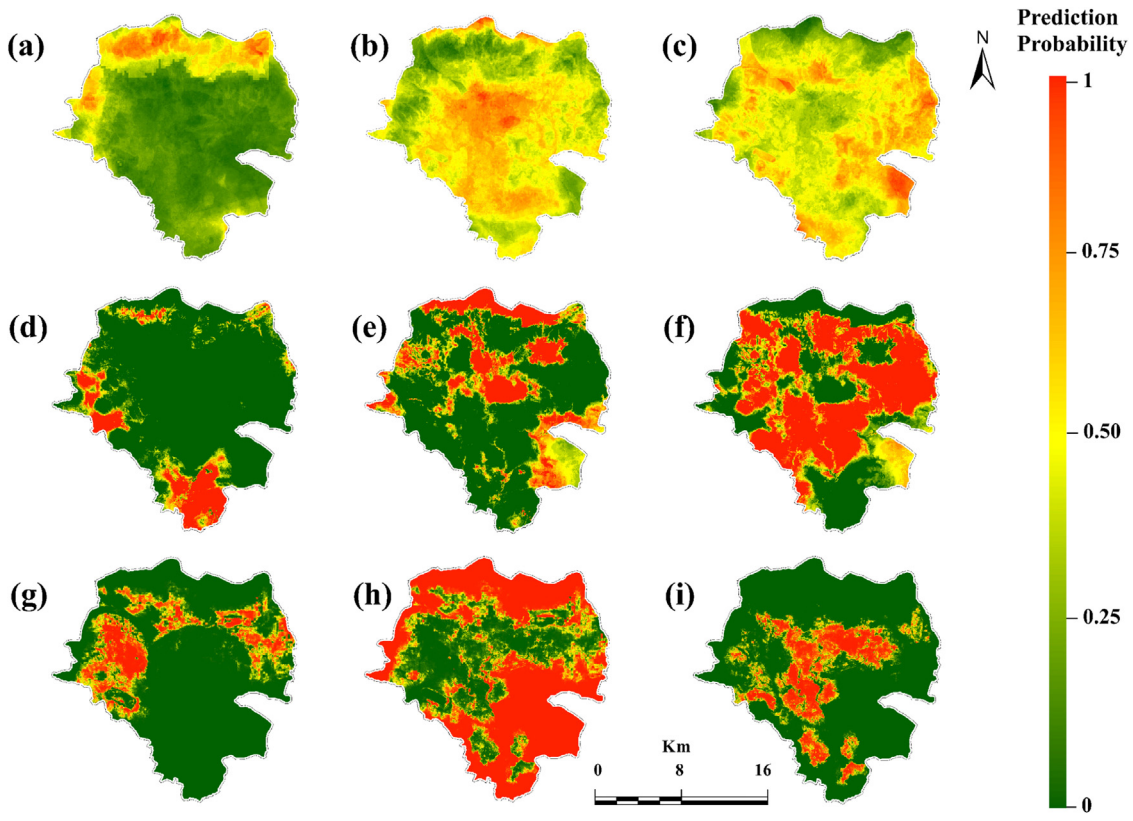

Fig. 13. Spatial distributions of PM_{10} prediction probabilities across the three MLAs. The top row represents RF classifications: (a) Good, (b) Moderate, and (c) UHealSsen categories. The middle row shows KNN results: (d) Good, (e) Moderate, and (f) UHealSsen categories. The bottom row illustrates NB predictions: (g) Good, (h) Moderate, and (i) UHealSsen categories. The colour gradient from green (low probability) to red (high probability) indicates the predicted likelihood of each category.

KNN, while demonstrating competitive performance, exhibited slightly lower sensitivity and specificity in the Moderate and UHealSesn classes, with a minor decline in balanced accuracy (92% vs. 94% for RF) (Fig. 12). Notably, the KNN model's over-reliance on neighbourhood-based classification techniques makes it susceptible to localized variations, which may lead to misclassifications, particularly in transitional zones between pollution levels [32]. The NB model performed the weakest among the three classifiers, particularly in the Moderate class, where balanced accuracy dropped to 0.89, and AUC to 0.96 (Fig. 4). The assumption of feature independence in NB may not hold for air pollution data, where strong interdependencies exist between meteorological factors, emission sources, and topographic influences [41]. Therefore, NB's spatial classifications appeared more generalized, reducing accuracy in identifying precise pollution hotspots.

Despite RF's superior performance, it is imperative to acknowledge its limitations. While practical, the ensemble approach of RF introduces substantial computational complexity compared to simpler models like KNN and NB. This computational demand may pose challenges for real-time implementation in resource-constrained monitoring systems. Additionally, the black-box nature of RF also complicates interpretability, potentially obscuring the relative influence of specific environmental variables on PM_{10} predictions [42–44].

The spatial distributions of PM_{10} prediction probabilities across the three classifiers (Fig. 13) provide further insights into their geospatial accuracy. The top row (a–c) presents RF's classifications for the Good, Moderate, and UHealSesn categories, demonstrating a highly detailed and spatially distinct prediction pattern with a clear delineation of pollution hotspots. The middle row (d–f) illustrates KNN's results, which capture trends similar to RF but with slightly reduced spatial precision. The bottom row (g–i) represents NB's predictions, which appear more generalized, missing the fine-scale detail observed in RF and KNN. RF's high specificity ensures accurate delineation of low-risk areas, while its strong sensitivity enables reliable identification of pollution hotspots [2,41]. The model's better handling of spatial autocorrelation makes it particularly suited for urban environments where pollution sources vary widely from one location to another. Fig. 13 (d-f) reveals that the KNN spatial distribution map shows more homogeneous classifications, which may result in less precise hotspot identification. Meanwhile, NB's spatial predictions (Fig. 13, g-i) demonstrate over-smoothing, reducing classification granularity and misidentifying certain pollution-prone areas [41].

Our model's results align with existing research on ML applications in air pollution modelling. AlThuwaynee et al. [2] and Tella et al. [30,45] have demonstrated RF superior predictive ability for PM_{10} classification, attributing its success to robust ensemble

learning techniques. Ahmad et al. [46] also found that RF outperforms traditional land-use regression models, particularly in urban settings with heterogeneous pollution sources. Moreover, Bozdağ et al. [7] highlighted RF's ability to handle high-dimensional environmental data, further reinforcing its suitability for PM_{10} spatial prediction. In contrast, Pant et al. [28] suggest that NB performs optimally in pollution forecasting due to its realistic independence assumptions. Our study's findings are consistent with these observations, confirming that RF provides the most reliable predictions for PM_{10} pollution in Addis Ababa.

The superior predictive performance of RF underscores its practical relevance for urban air quality monitoring and environmental policy planning [47]. Given the serious health risks associated with PM_{10} exposure, integrating RF-based models into real-time monitoring systems can significantly enhance early warning mechanisms. Deploying RF-based predictive systems at strategic locations throughout Addis Ababa could substantially augment the city's existing air quality monitoring infrastructure, enabling preemptive interventions during forecasted pollution events. Specifically, RF predictions could inform dynamic traffic management systems, activating congestion mitigation protocols when elevated PM_{10} concentrations are anticipated in densely populated corridors. Additionally, industrial emission controls could be temporally adjusted based on RF forecasts, with more stringent limitations imposed during predicted pollution peaks. The spatial precision of RF models could further guide urban greening initiatives, prioritizing vegetation barriers in consistently identified pollution hotspots.

RF prediction models also enable public health authorities to issue targeted pollution warnings to vulnerable populations when UHealSesn conditions are anticipated, while integration with cellular networks facilitates personalized air quality alerts that empower residents to adjust activities based on hyperlocal forecasts. At the policy level, RF-generated spatial predictions support evidence-based regulatory frameworks, enabling zone-specific emission standards tailored to local vulnerability and dispersion patterns, significantly improving uniform regulations that inadequately address Addis Ababa's complex topography and urban structure. This empirical study demonstrates RF's superiority over KNN and NB models for PM_{10} prediction in Addis Ababa, confirming its robust capability to capture spatial pollution variations with statistical significance. Despite computational complexities, RF's exceptional predictive accuracy renders it optimal for urban air quality monitoring. Implementation in real-time systems could substantially enhance pollution forecasting and mitigation strategies. Future research should explore hybrid deep-learning approaches to improve predictive accuracy while addressing computational limitations. Meticulous data preprocessing and methodological optimization proved critical for model performance, underscoring the importance of rigorous analytical frameworks for operational implementation in complex urban environments [2,6,7].

Limitations

This study used MLAs, including RF, KNN, and NB, which have certain limitations. These algorithms can identify statistical patterns in data but cannot provide insights into the underlying processes that lead to air pollution [2]. The quality of the training data affects the accuracy of these algorithms, and any noise or unaccounted-for variables can negatively impact their performance. The NB algorithm assumes feature independence, which may not hold in real-world scenarios, resulting in suboptimal performance [14]. Although the study used evaluation metric techniques, the performance of the models in real-world applications can be affected by factors not captured in the training data, such as changes in emission patterns or urban development. Despite these limitations, this study provides a valuable foundation for future research that should integrate advanced techniques and additional data sources to improve the robustness and interpretability of air pollution prediction models.

Notably, while these algorithms are established methods, their application and optimization in the context of PM_{10} prediction represent a novel approach. Our study is based on carefully selecting and fine-tuning these algorithms specifically for the spatial Prediction of PM_{10} levels. Additionally, we conducted extensive model validations to demonstrate the effectiveness of these algorithms in this specific application. The detailed CV procedures, hyperparameter optimization, and comparative analysis of different parameter values provide solid experimental support for our results. These rigorous experimental processes validate the algorithms' effectiveness and provide valuable insights into their performance in the specific context of PM_{10} prediction.

Ethics statements

This work does not use human subjects, animals, or data collected through social media as research materials.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Kalid Hassen Yasin: Methodology, Software, Formal analysis, Investigation, Resources, Data curation, Writing – original draft. **Muaz Ismael Yasin:** Methodology, Resources, Writing – review & editing. **Anteneh Derribew Iguala:** Resources, Writing – review & editing. **Tadele Bedo Gelete:** Methodology, Writing – review & editing. **Erana Kebede:** Resources, Writing – review & editing.

Data availability

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.12825036.

Acknowledgments

The authors express their gratitude to all the data providers referenced in the article for providing the necessary data for this analysis.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] D. Vilcins, R.C. Christofferson, J.-H. Yoon, S.N. Nazli, P.D. Sly, S.A. Cormier, G. Shen, Updates in Air Pollution: Current Research and Future Challenges, Ann. Glob. Heal. 90 (2024), doi:10.5334/aogh.4363.
[2] O.F. AlThuwaynee, S.W. Kim, M.A. Najemaden, A. Aydda, A.L. Balogun, M.M. Fayyadh, H.J. Park, Demystifying uncertainty in PM_{10} susceptibility mapping using variable drop-off in extreme-gradient boosting (XGB) and random forest (RF) algorithms, Environ. Sci. Pollut. Res. 28 (2021) 43544–43566, doi:10.1007/s11356-021-13255-4.
[3] X. Ren, Z. Mi, P.G. Georgopoulos, Comparison of Machine Learning and Land Use Regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States, Environ. Int. 142 (2020) 105827, doi:10.1016/j.envint.2020.105827.
[4] W.N. Shaziayani, A.Z. Ul-Saufie, H. Ahmat, D. Al-Jumeily, Coupling of quantile regression into boosted regression trees (BRT) technique in forecasting emission model of PM_{10} concentration, Air Qual. Atmos. Heal. 14 (2021) 1647–1663, doi:10.1007/s11869-021-01045-3.
[5] N.v. Tkachenko, Optical Spectroscopy, Elsevier, 2006, doi:10.1016/B978-0-444-52126-2.X5024-2.
[6] L. Lin, Y. Liang, L. Liu, Y. Zhang, D. Xie, F. Yin, T. Ashraf, Estimating PM2.5 Concentrations Using the Machine Learning RF-XGBoost Model in Guanzhong Urban Agglomeration, China, Remote Sens. 14 (2022) 5239, doi:10.3390/rs14205239.
[7] A. Bozdağ, Y. Dokuz, Ö.B. Gökçek, Spatial prediction of PM_{10} concentration using machine learning algorithms in Ankara, Turkey, Environ. Pollut. 263 (2020) 114635, doi:10.1016/j.envpol.2020.114635.
[8] L. Davidson, K. Kline, S. Klein, K. Windisch, The Normalization Process, in: Pro SQL Serv. 2008 Relational Database Des. Implement., Apress, Berkeley, CA, 2008, pp. 117–175, doi:10.1007/978-1-4302-0867-9_4.
[9] K.H. Yasin, M.I. Yasin, A.D. Iguala, T.B. Gelete, Dataset of PM_{10} Level and Predictors in Addis Ababa, (2024). https://doi.org/10.5281/zenodo.12825036.
[10] N. Kumar, F.M. Hamzah, M. Diantoro, N.A. Muhd Zailani, Suman, Physiochemical characterization of ambient PM_{10} and PM2.5 in an urban environment, Curr. Appl. Phys. 71 (2025) 57–69, doi:10.1016/j.cap.2024.12.006.
[11] G. Raheja, J. Nimo, E.K.-E. Appoh, B. Essien, M. Sunu, J. Nyante, M. Amegah, R. Quansah, R.E. Arku, S.L. Penn, M.R. Giordano, Z. Zheng, D. Jack, S. Chillrud, K. Amegah, R. Subramanian, R. Pinder, E. Appah-Sampong, E.N. Tetteh, M.A. Borketey, A.F. Hughes, D.M. Westervelt, Low-Cost Sensor Performance Intercomparison, Correction Factor Development, and 2+ Years of Ambient PM 2.5 Monitoring in Accra, Ghana, Environ. Sci. Technol. 57 (2023) 10708–10720, doi:10.1021/acs.est.2c09264.
[12] E.N. Mekonnen, A. Fetene, E. Gebremariam, Grid-based climate variability analysis of Addis Ababa, Ethiopia, Heliyon 10 (2024) e27116, doi:10.1016/j.heliyon.2024.e27116.
[13] A.T. Weldegebriel, M. Tekalign, A. Van Rompaey, Socio-spatial analysis of regime shifts in Addis Ababa's urbanisation, Appl. Geogr. 154 (2023) 102918, doi:10.1016/j.apgeog.2023.102918.
[14] M. Imam, S. Adam, S. Dev, N. Nesa, Air quality monitoring using statistical learning models for sustainable environment, Intell. Syst. with Appl. 22 (2024) 200333, doi:10.1016/j.iswa.2024.200333.
[15] K.H. Yasin, A.D. Iguala, T.B. Gelete, Spatiotemporal analysis of urban expansion and its impact on farmlands in the central Ethiopia metropolitan area, Discov. Sustain. 6 (2025) 36, doi:10.1007/s43621-024-00749-7.
[16] A. Baruah, D. Bousiotis, S. Damayanti, A. Bigi, G. Ghermandi, O. Ghaffarpasand, R.M. Harrison, F.D. Pope, A novel spatiotemporal prediction approach to fill air pollution data gaps using mobile sensors, machine learning and citizen science techniques, Npj Clim. Atmos. Sci. 7 (2024) 310, doi:10.1038/s41612-024-00859-z.
[17] B. Choubin, M. Abdolshahnejad, E. Moradi, X. Querol, A. Mosavi, S. Shamshirband, P. Ghamisi, Spatial hazard assessment of the PM_{10} using machine learning models in Barcelona, Spain, Sci. Total Environ. 701 (2020) 2–11, doi:10.1016/j.scitotenv.2019.134474.
[18] D.J. Kleine, R. Zalakeviciute, M. Gonzalez, Y. Rybarczyk, Modeling PM 2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters, J. Electr. Comput. Eng. 2017 (2017) 1–14, doi:10.1155/2017/5106045.
[19] A.B. Shiferaw, A. Kumie, W. Tefera, The spatial and temporal variation of fine particulate matter pollution in Ethiopia: Data from the Atmospheric Composition Analysis Group (1998–2019), PLoS One 18 (2023) e0283457, doi:10.1371/journal.pone.0283457.
[20] A.N. Amare, S. Sorsa, Z. Gebremariam, Levels and health risk assessments of particulate matter and inorganic gaseous pollutants in urban and industrial areas of Hawassa city, Ethiopia, Heliyon 10 (2024) e33286, doi:10.1016/j.heliyon.2024.e33286.
[21] T.R. Govindasamy, N. Chetty, Machine learning models to quantify the influence of PM_{10} aerosol concentration on global solar radiation prediction in South Africa, Clean. Eng. Technol. 2 (2021) 100042, doi:10.1016/j.clet.2021.100042.
[22] K. Okorn, L.T. Iraci, An overview of outdoor low-cost gas-phase air quality sensor deployments: current efforts, trends, and limitations, Atmos. Meas. Tech. 17 (2024) 6425–6457, doi:10.5194/amt-17-6425-2024.
[23] T.D. Morapedi, I.C. Obagbuwa, Air pollution particulate matter (PM2.5) prediction in South African cities using machine learning techniques, Front, Artif. Intell. 6 (2023), doi:10.3389/frai.2023.1230087.
[24] F. Carotenuto, A. Bisignano, L. Brilli, G. Gualtieri, L. Giovannini, Low-cost air quality monitoring networks for long-term field campaigns: A review, Meteorol. Appl. 30 (2023), doi:10.1002/met.2161.
[25] A. Houdou, K. Khomsi, L.D. Monache, W. Hu, S. Boutayeb, L. Belyamani, F. Abdulla, W.K. Al-Delaimy, M. Khalis, Predicting Particulate Matter (PM_{10}) Levels in Morocco: A 5-Day Forecast Using the Analog Ensemble Method., (2024). https://doi.org/10.21203/rs.3.rs-4619478/v1.
[26] S.N. Jida, H. Jean-François, P. Chesse, Artificial Neural Network Modelling to Predict PM2.5 and PM_{10} Exhaust Emissions from On-Road Vehicles in Addis Ababa, Ethiopia, Environ. Pollut. Clim. Chang. 6 (2022) 279 Artificial Neural Network Modelling to Predict PM2.5 and PM_{10} Exhaust Emissions from On-Road Vehicles in Addis Ababa, Ethiopia.
[27] L. Breiman, A. Cutler, Breiman and Cutler's Random Forests for Classification and Regression, 2022. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf.
[28] A. Pant, K. Pant, N. Pathak, M. Ram, Prediction of Particulate Matter (PM2.5) for Industrial Area Based on Naive Bayes Classifier, in: Proc. Data Anal. Manag. ICDAM 2023. Lect. Notes Networks Syst., 2024, pp. 189–195, doi:10.1007/978-981-99-6547-2_15.
[29] F. Alzu'bi, A. Al-Rawabdeh, A. Almagbile, Predicting air quality using random forest: A case study in Amman-Zarqa, Egypt. J. Remote Sens. Sp. Sci. 27 (2024) 604–613, doi:10.1016/j.ejrs.2024.07.004.
[30] A. Tella, A.-L. Balogun, GIS-based air quality modelling: spatial prediction of PM_{10} for Selangor State, Malaysia using machine learning algorithms, Environ. Sci. Pollut. Res. 29 (2022) 86109–86125, doi:10.1007/s11356-021-16150-0.
[31] F. Mohammadi, H. Teiri, Y. Hajizadeh, A. Abdolahnejad, A. Ebrahimi, Prediction of atmospheric PM2.5 level by machine learning techniques in Isfahan, Iran, Sci. Rep. 14 (2024) 2109, doi:10.1038/s41598-024-52617-z.
[32] B. Chao, H.Guang Qiu, Air pollution concentration fuzzy evaluation based on evidence theory and the K-nearest neighbor algorithm, Front. Environ. Sci. 12 (2024), doi:10.3389/fenvs.2024.1243962.

[33] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Leonardo 7 (1974) 370, doi:10.2307/1573081.
[34] Z. Yao, W.L. Ruzzo, A Regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data, BMC Bioinformatics 7 (2006) S11, doi:10.1186/1471-2105-7-S1-S11.
[35] K.C. Atmakuri, K.V Prasad, Urban Air Quality Analysis And Aqi Prediction Using Improved Knn Classifier, J. Pharm. Negat. Results ¦. 13 (2023) 2022, doi:10.47750/pnr.2022.13.S07.854.
[36] H. Shahabi, A. Shirzadi, K. Ghaderi, E. Omidvar, N. Al-Ansari, J.J. Clague, M. Geertsema, K. Khosravi, A. Amini, S. Bahrami, O. Rahmati, K. Habibi, A. Mohammadi, H. Nguyen, A.M. Melesse, B. Bin Ahmad, A. Ahmad, Flood Detection and Susceptibility Mapping Using Sentinel-1 Remote Sensing Data and a Machine Learning Approach: Hybrid Intelligence of Bagging Ensemble Based on K-Nearest Neighbor Classifier, Remote Sens. 12 (2020) 266, doi:10.3390/rs12020266.
[37] I. Rish, An empirical study of the naive bayes classifier, in: IJCAI 2001 Work. Empir. Methods Artif. Intell, 2001, pp. 41–46.
[38] R.W. Gore, D.S. Deshpande, An approach for classification of health risks based on air quality levels, in: 2017 1st Int. Conf. Intell. Syst. Inf. Manag, IEEE, 2017, pp. 58–61, doi:10.1109/ICISIM.2017.8122148.
[39] M.J. Sairam, V. Nagaraju, Air pollution hotspot identification to prevent post effects of pollution using naive bayes over random forest, Int. Conf. Adv. Des. Dev. Eng. Process. Charact. ADDEPC 2021, 2024, doi:10.1063/5.0197506.
[40] H. Wang, H. Wang, Z. Wu, Y. Zhou, Using Multi-Factor Analysis to Predict Urban Flood Depth Based on Naive Bayes, Water 13 (2021) 432, doi:10.3390/w13040432.
[41] M.M. Rahman, M.E.H. Nayeem, M.S. Ahmed, K.A. Tanha, M.S.A. Sakib, K.M.M. Uddin, H.M.H. Babu, AirNet: predictive machine learning model for air quality forecasting using web interface, Environ. Syst. Res. 13 (2024) 44, doi:10.1186/s40068-024-00378-z.
[42] M. Aria, C. Cuccurullo, A. Gnasso, A comparison among interpretative proposals for Random Forests, Mach. Learn. with Appl. 6 (2021) 100094, doi:10.1016/j.mlwa.2021.100094.
[43] K. Moulaei, M.R. Afrash, M. Parvin, S. Shadnia, M. Rahimi, B. Mostafazadeh, P.E.T. Evini, B. Sabet, S.M. Vahabi, A. Soheili, M. Fathy, A. Kazemi, S. Khani, S.M. Mortazavi, S.M. Hosseini, Explainable artificial intelligence (XAI) for predicting the need for intubation in methanol-poisoned patients: a study comparing deep and machine learning models, Sci. Rep. 14 (2024) 15751, doi:10.1038/s41598-024-66481-4.
[44] W. Feng, C. Ma, G. Zhao, R. Zhang, FSRF:An Improved Random Forest for Classification, in: 2020 IEEE Int. Conf. Adv. Electr. Eng. Comput. Appl. AEECA), IEEE, 2020, pp. 173–178, doi:10.1109/AEECA49918.2020.9213456.
[45] A. Tella, A.-L. Balogun, N. Adebisi, S. Abdullah, Spatial assessment of PM_{10} hotspots using Random Forest, K-Nearest Neighbour and Naïve Bayes, Atmos. Pollut. Res. 12 (2021) 1–12, doi:10.1016/j.apr.2021.101202.
[46] N. Ahmad, A.Z. Ul-Saufie, W.N. Shaziayani, A.W. Zainan Abidin, N.E. Sharmila Zulazm, S. Harb, Evaluating the Performance of Random Forest and Multiple Linear Regression for Higher Observed PM_{10} Concentrations, Israa Univ, J. Appl. Sci. 6 (2022) 72–90, doi:10.52865/WHPM9019.
[47] L. Mamić, M. Gašparović, G. Kaplan, Developing PM2.5 and PM_{10} prediction models on a national and regional scale using open-source remote sensing data, Environ. Monit. Assess. 195 (2023) 644, doi:10.1007/s10661-023-11212-x.