Journal of Chemistry and Allied Sciences
Volume 2, Issue 1 | 2026 | pp. 139~149
Original Research Article | ***Available online at*** **https://jcasjournal.org/**

MACHINE LEARNING PREDICTION OF INDOOR $PM_{2.5}$-BOUND DIBENZO[a,h]ANTHRACENE IN CHILDREN'S CHURCH FACILITIES IN SOUTHERN NIGERIA USING MICROCLIMATIC DATA

Ukeme Donatus Archibong*[1], Christopher Ukuegboho Michael[2]

[1]Department of Applied Chemical Science Laboratory Technology, Faculty of Science Laboratory Technology, University of Benin, Benin City, Edo State, Nigeria

[2]Department of Science Laboratory Technology, Delta State Polytechnic, Otefe Oghara, Ethiope West, Delta State, Nigeria

Article History*: Received January 2026; Revised January 2026; Accepted February 2026; Published online February 2026*

Corresponding author*:* *Ukeme Donatus Archibong* (ukeme.archibong@uniben.edu ***ORCID****: https://orcid.org/0009-0001-5949 6640;* ***Tel****: +234 802 373 5803)*

Article Information

Citation: **Citation**: Archibong, U. D. & Michael, U. C. (2026). Machine Learning Prediction of Indoor $PM_{2.5}$-Bound Dibenzo[a,h]anthracene in Children's Church Facilities in Southern Nigeria Using Microclimatic Data. Journal of Chemical and Allied Science, 2(1), 139-149.
DOI:
https://doi.org/10.60787/jcas.vol2no1.1

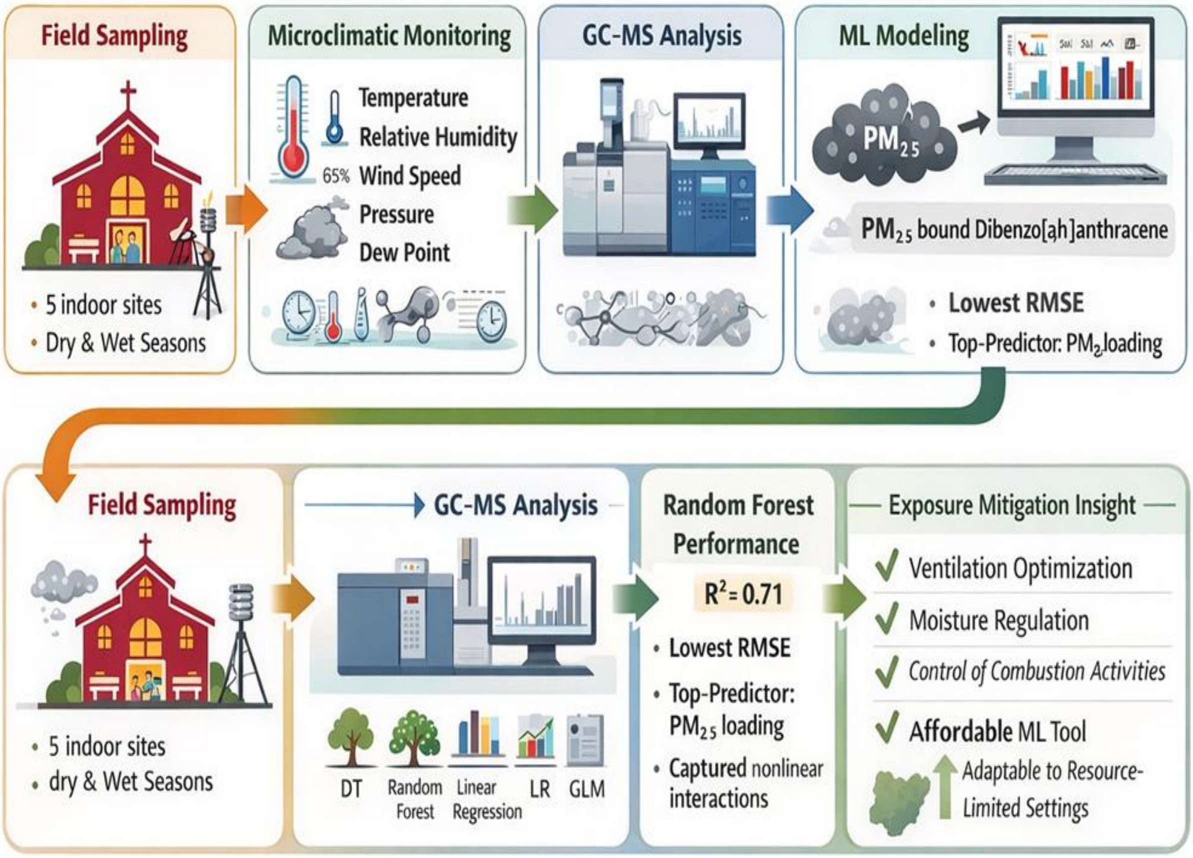
The Official Publication of the Tropical Research and Allied Network (TRANet), Department of Chemistry, *Federal University of Technology, Minna.*

Abstract

Polycyclic aromatic hydrocarbons (PAHs) are hazardous semivolatile organic compounds frequently detected in indoor environments, several of which are designated by the United States Environmental Protection Agency (USEPA) as priority pollutants due to their carcinogenic potential. Dibenzo[a,h]anthracene (DahA), a high-molecular-weight PAH strongly associated with combustion-derived particulate matter, is classified by the International Agency for Research on Cancer (IARC) as probably carcinogenic to humans. This study applied supervised machine learning (ML) to predict indoor concentrations of $PM_{2.5}$-bound DahA in children's church facilities in Ugbowo, southern Nigeria, using microclimatic predictors. A total of 30 indoor air samples were collected from 5 locations across dry and wet seasons, and concentrations of the 16 USEPA priority PAHs were quantified by gas chromatography–mass spectrometry. Concurrent measurements of temperature, relative humidity, wind speed, atmospheric pressure, and dew point were obtained. Decision tree (DT), random forest (RF), linear regression (LR), and generalized linear model (GLM) algorithms were trained and evaluated using mean squared error (MSE), root mean squared error (RMSE), and the coefficient of determination (R^2). Random Forest outperformed the other models ($R^2 = 0.71$), demonstrating superior capability in capturing nonlinear interactions between $PM_{2.5}$ loading and microclimatic variables. While DahA prediction showed greater variability due to relatively low ambient concentrations, the Random Forest model effectively reproduced temporal and microenvironmental trends in DahA concentrations. The results demonstrate the utility of ensemble ML approaches for indoor PAH assessment and highlight the relevance of microclimatic controls for exposure mitigation in child-centric indoor environments.

Keywords: polycyclic aromatic hydrocarbons; indoor air quality; dibenzo[a,h]anthracene; $PM_{2.5}$; machine learning

Graphical Abstract



1.0 INTRODUCTION

Polycyclic aromatic hydrocarbons (PAHs) are ubiquitous products of incomplete combustion of organic matter, including biomass, fossil fuels, tobacco, incense, and candles. Owing to their semivolatile nature, PAHs partition between the gas phase and particulate matter, with high-molecular-weight congeners preferentially associated with fine particles ($PM_{2.5}$), thereby enhancing inhalation exposure and deep lung deposition [1]. The USEPA has identified 16 priority PAHs based on their toxicity and prevalence in the environment, several of which are established or probable human carcinogens [2].

Dibenzo[a,h]anthracene (DahA) is among the most toxic PAHs, with a toxic equivalency factor comparable to benzo[a]pyrene and strong evidence of genotoxicity following metabolic activation [3,4,5]. Indoor exposure to particle-bound carcinogenic PAHs is of particular concern in child-centric environments, where children's physiological vulnerability and behavioural patterns (e.g., proximity to the floor and hand-to-mouth activity) may increase effective dose. The indoor worship facilities are a microenvironment that has not been effectively studied because of the presence of activities like candle and incense burning, with minimal ventilation that can increase the burden of PAH indoors.

Children's worship facilities are an understudied indoor microenvironment where combustion-related activities, such as candle and incense burning, can elevate PAH levels, particularly under poor ventilation conditions [6]. Additional contributors to $PM_{2.5}$-bound dibenzo[a,h]anthracene (DahA) in this microenvironment may include outdoor traffic emission infiltration, resuspended contaminated dust, and thermal degradation of building materials in densely occupied, low air-exchange spaces [7].

Meteorological and microclimatic conditions regulate the dynamics of indoor PAH by affecting the level of ventilation, particle suspension, and deposition. Gas-particle partitioning and hygroscopic growth are influenced by temperature and humidity, whereas the wind speed and pressure are used as proxies of ventilation and indoor atmospheric stability. Traditional statistical models do not adequately model nonlinear and interactive effects of these predictors. On the other hand, machine learning (ML) methods, particularly ensemble tree-based algorithms such as random forests, have shown robust performance in air pollution modeling by accommodating nonlinearities and multicollinearity between predictors effectively [8,9].

To the best of our knowledge, this is the first study to apply supervised machine learning to predict $PM_{2.5}$-bound dibenzo[a,h]anthracene concentrations in children's worship facilities within a low- and middle-income setting. Unlike previous PAH studies that emphasize concentration assessment alone, this work integrates field-based microclimatic monitoring with ensemble machine-learning modeling to identify key environmental drivers of carcinogenic PAH variability and provide a predictive exposure-assessment framework applicable in resource-limited environments. This work addresses this research gap by (i) quantifying indoor $PM_{2.5}$-bound DahA across seasons, (ii) developing and comparing ML models utilizing microclimatic variables, and (iii) estimating model performance through standard regression metrics. The overall goal is to offer a data-based platform of exposure measurement and alleviation in susceptible indoor micro-conditions.

2.0 MATERIALS AND METHODS

2.1 Study Area

The current study was conducted on five sampled children church centers along the Ugbowo axis of the Benin City, Edo State, Southern Nigeria, and the goal was to measure the parameters of indoor air quality and the degree of anthropogenic impact. Ugbowo is in the middle of the Benin City that is situated between 6°16'N and 6°26'N and longitudes 5°32'E and 5°38'E and is under the humid tropical climate zone with two different seasons: the wet and the dry season. Five church facilities were sampled, and measurements were conducted accordingly, resulting in 5 distinct indoor sampling locations. Field sites were coded in a systematic way and the exact geographical coordinates along with the altitude of each sampling location is thoroughly described in Table 1. Table 1 has the coded names of the sampling sites, geographic positions and altitude of the sites.

Table 1: Geographic Coordinates of Indoor Air Sampling Locations

S/N	Location (Coded)	Latitude (Decimal Degrees)	Longitude (Decimal Degrees)	Elevation (Meters)
1	ASC	6.4008	5.6119	75.8
2	SAC	6.4016	5.6118	82.1
3	SPC	6.3744	5.6134	85.4
4	WCU	6.3876	5.6190	80.5
5	HRC	6.2155	5.6127	72.4

2.2 Sampling Design

Indoor air sampling was conducted at 5 locations within children's church facilities within Ugbowo, southern Nigeria. Triplicate samples were collected per site across dry and wet seasons, yielding 30 valid samples after quality control. Each sampling event represented an 8-h integrated indoor exposure under typical occupancy conditions collected at a height of 1.5-2.0 meters above the ground within the air breathing zone using the Apex2IS Cassella standard pump, coupled with a conical inhalable sampling (CIS) head at the flow rate of 3.5 litres per minute (LPM). Samples were collected using the standard gravimetric method [2].

2.3 Microclimatic Data

Temperature (°C), relative humidity (%), wind speed (m s^{-1}), atmospheric pressure (hPa), and dew point (°C) were recorded concurrently with sampling at 5-minute intervals using an on-site automatic weather monitoring system (professional weather station) to determine influences on pollutant dynamics. Daily means were used as predictors in ML models. $PM_{2.5}$ mass concentration was also included as a predictor variable, given its role as the carrier phase for particle-bound DahA [2].

2.4 PAH Analysis

Particulate-bound PAHs were collected on pre-baked quartz fibre filters, while gas-phase PAHs were trapped using polyurethane foam plugs. Filters were ultrasonically extracted using dichloromethane:hexane (1:1 v/v) for 30 min. Extracts were concentrated and cleaned using silica gel column chromatography before GC–MS analysis. The 16 USEPA priority PAHs were quantified by gas chromatography–mass spectrometry following established extraction and cleanup protocols. Gas chromatography–mass spectrometry (GC–MS) analysis was performed using an HP-5MS capillary column (30 m × 0.25 mm internal diameter × 0.25 µm film thickness) with helium employed as the carrier gas. The oven temperature was initially held at 60 °C for 2 minutes and subsequently ramped to 300 °C at a rate of 10 °C min^{-1}. The method detection limit (LOD) for dibenzo[a,h]anthracene (DahA) was 0.01 ng m^{-3}, while the limit of quantification (LOQ) was 0.03 ng m^{-3}. Surrogate recovery efficiencies ranged from 82% to 105%, indicating acceptable analytical performance and method reliability (Archibong *et al.,* 2025). DahA was prioritized in subsequent analyses due to its high toxic potency.

2.5 Machine Learning Framework

The four supervised regression algorithms used were decision tree, random forest, linear regression and generalized linear model. Python version 3.12 and scientific libraries NumPy version 2.0, pandas version 2.2, scikit-learn version 1.5, SciPy version 1.13 and matplotlib version 3.9 were used to perform data preprocessing, exploratory data analysis, multivariate statistical assessments and machine-learning modeling. $PM_{2.5}$-bound dibenzo[a,h]anthracene (DahA) concentration was used as the target variable, while $PM_{2.5}$ mass, temperature, relative humidity, wind speed, atmospheric pressure, and dew point served as predictors. All variables were standardized prior to modelling. The dataset was divided into training and testing sets using an 80:20 split, and model robustness was evaluated through 5-fold cross-validation. Hyperparameters were optimized via grid search; for the Random Forest model, tuning parameters included the number of trees (100–500), maximum tree depth (5–20), and minimum samples required for node splitting (2–10). Model performance was assessed using R^2, RMSE, and MAE. Also, to assess multicollinearity among predictor variables before model development, Variance Inflation Factor (VIF) analysis was performed. Each independent variable was regressed against all other predictors, and the resulting R^2 values were used to calculate the VIF for each variable as represented in equation (1)

$$VIF_i = \frac{1}{1- R_i^2} \qquad (1)$$

Where:

VIF_i = Variance Inflation Factor for the i^{th} predictor variable X_i, R_i^2 = the coefficient of determination when X_i is regressed on all other independent variables in the model and $1 - R_i^2$ = the proportion of X_i's variance that is not explained by the other predictors.

Variables with VIF values exceeding 5 were considered to exhibit high multicollinearity and were carefully evaluated before inclusion in the regression models. This procedure ensured stability and reliability of the estimated regression coefficients [10].

3.0 Results and Discussion

Table 2. Monthly and seasonal descriptive statistics (Mean ± SD) of $PM_{2.5}$-bound dibenzo[a,h]anthracene and microclimatic parameters across indoor sampling sites.

Month/Season	DahA (ng m⁻³)	$PM_{2.5}$ (µg m⁻³)	TEMP (°C)	RH (%)	WS (m s⁻¹)	PRESS (hPa)	DP (°C)
January	3.62 ± 1.84	46.8 ± 19.5	26.9 ± 2.1	69.2 ± 8.1	1.12 ± 0.48	1012.4 ± 3.1	20.1 ± 1.9
February	3.28 ± 1.71	44.1 ± 17.9	27.6 ± 2.0	66.8 ± 7.6	1.20 ± 0.51	1011.8 ± 2.9	20.6 ± 2.0
March	2.91 ± 1.46	39.7 ± 16.3	28.8 ± 2.3	70.5 ± 8.9	1.35 ± 0.62	1010.2 ± 3.0	21.8 ± 2.1
April	2.47 ± 1.29	36.9 ± 14.8	29.6 ± 2.4	73.1 ± 9.2	1.44 ± 0.66	1009.4 ± 2.8	22.7 ± 2.0
May	2.21 ± 1.14	33.8 ± 13.6	29.1 ± 2.2	75.4 ± 8.7	1.58 ± 0.71	1008.9 ± 2.7	23.4 ± 1.9
June	2.03 ± 1.05	31.6 ± 12.9	27.8 ± 1.9	78.6 ± 7.9	1.66 ± 0.73	1008.2 ± 2.6	24.1 ± 1.8
July	1.89 ± 0.98	29.4 ± 12.1	26.7 ± 1.7	81.2 ± 6.8	1.74 ± 0.75	1007.6 ± 2.5	24.5 ± 1.7
August	1.94 ± 1.02	30.2 ± 12.7	26.5 ± 1.6	82.1 ± 6.5	1.71 ± 0.72	1007.4 ± 2.6	24.6 ± 1.6
September	2.18 ± 1.16	33.1 ± 13.8	27.2 ± 1.8	79.4 ± 7.2	1.59 ± 0.69	1008.1 ± 2.7	24.0 ± 1.8
October	2.63 ± 1.33	36.5 ± 15.2	28.1 ± 2.0	75.3 ± 8.0	1.42 ± 0.63	1009.3 ± 2.8	23.1 ± 1.9
November	3.05 ± 1.58	41.2 ± 17.1	28.6 ± 2.1	71.6 ± 7.9	1.26 ± 0.55	1010.8 ± 2.9	21.9 ± 2.0
December	3.44 ± 1.76	45.9 ± 18.8	27.4 ± 2.2	68.7 ± 8.4	1.15 ± 0.50	1012.0 ± 3.0	20.5 ± 2.1
Dry Season	3.45 ± 1.77	45.6 ± 18.7	27.3 ± 2.1	68.2 ± 8.0	1.16 ± 0.50	1011.2 ± 3.0	21.4 ± 2.2
Wet Season	1.95 ± 1.02	30.4 ± 12.6	27.0 ± 1.7	80.6 ± 6.9	1.70 ± 0.74	1008.6 ± 2.8	23.8 ± 2.1

3.1 Seasonal Variability of DahA and Microclimate Variables

Descriptive statistics of the concentration of $PM_{2.5}$-bound dibenzo[a,h]anthracene (DahA) and the related microclimatic parameters of monthly and seasonal data are summarized in Table 2. There was a high degree of seasonal variability in dahA, with the mean values of 1.89 ± 0.98 ng m⁻³ (July) and 3.62 ± 1.84 ng m⁻³ (January). Dry-season levels (3.45 ± 1.77 ng m⁻³) were approximately 1.8-fold higher than wet-season levels (1.95 ± 1.02 ng m⁻³), mirroring the seasonal pattern observed for $PM_{2.5}$ mass concentrations (45.6 ± 18.7 µg m⁻³ in the dry season versus 30.4 ± 12.6 µg m⁻³ in the wet season). $PM_{2.5}$ concentrations exceeded the World Health Organization (WHO) and the U.S. Environmental Protection Agency (USEPA) guideline values across both seasons, indicating persistently elevated indoor particulate burdens [11,12].

There were coherent seasonal variations in the microclimatic conditions, which explained the dynamics of pollutants observed. The wet season had a high relative humidity, dew point and wind speed, which favoured better ventilation, hygroscopic growth and fine particle deposition. On the other hand, the increase in atmospheric pressure and the relative decrease in air circulation during the dry season helped to build up the pollutants. The co-elevation of $PM_{2.5}$ and $PM_{2.5}$-bound DahA under dry-season conditions can thus be related to the synergistic effects of increased indoor combustion-based activities, low ventilation and low moisture-based scavenging. By contrast, the attenuation process experienced in the wet season is in line with increased elimination of particle-bound PAHs through moisture-related mechanisms [13,14]. This seasonal pattern aligns with previous West African observations of dry-season amplification of indoor PAHs [15] and underscores the importance of accounting for microclimatic variability in predictive modeling.

3.2 Multivariate Structure and Predictor Interrelationships

Table 3. Correlation Matrix showing Parameter Relationship between Variables

Variable	DahA	$PM_{2.5}$	TEMP	RH	WS	PRESS	DP
DahA	1.000						
$PM_{2.5}$	0.996	1.000					
TEMP	0.072	0.109	1.000				
RH	-0.947	-0.959	-0.316	1.000			
WS	-0.994	-0.995	-0.158	0.960	1.000		
PRESS	0.975	0.978	0.110	-0.949	-0.972	1.000	
DP	-0.973	-0.977	-0.124	0.956	0.969	-0.994	1.000

Pearson correlation analysis (Table 3) revealed strong coupling between DahA and $PM_{2.5}$ ($r = 0.996$), supporting the role of fine particles as the dominant carrier phase for high-molecular-weight PAHs. Variance Inflation Factor (VIF) analysis confirmed multicollinearity among certain meteorological predictors. However, Random Forest models are robust to multicollinearity due to their ensemble tree-based architecture. The strong coupling observed between $PM_{2.5}$ and dibenzo[a,h]anthracene in the present study is consistent with source-oriented investigations showing preferential enrichment of high-molecular-weight PAHs on fine particles and associated carcinogenic risk [16,17]. Relative humidity, wind speed, and dew point exhibited strong inverse relationships with DahA, reflecting the roles of ventilation and moisture-driven deposition in attenuating indoor particle-bound PAHs. The pronounced reduction in $PM_{2.5}$-bound DahA during the wet season is consistent with enhanced wet deposition and moisture-driven scavenging of particle-bound PAHs reported in urban atmospheres [14].

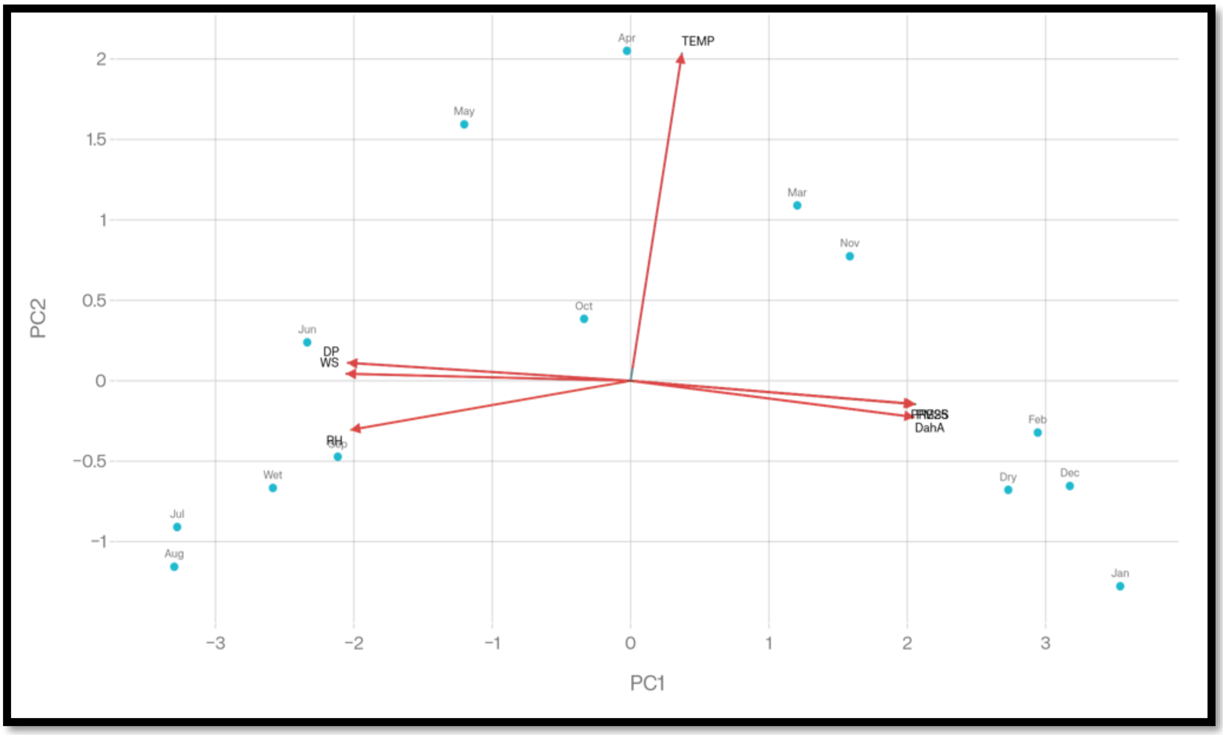
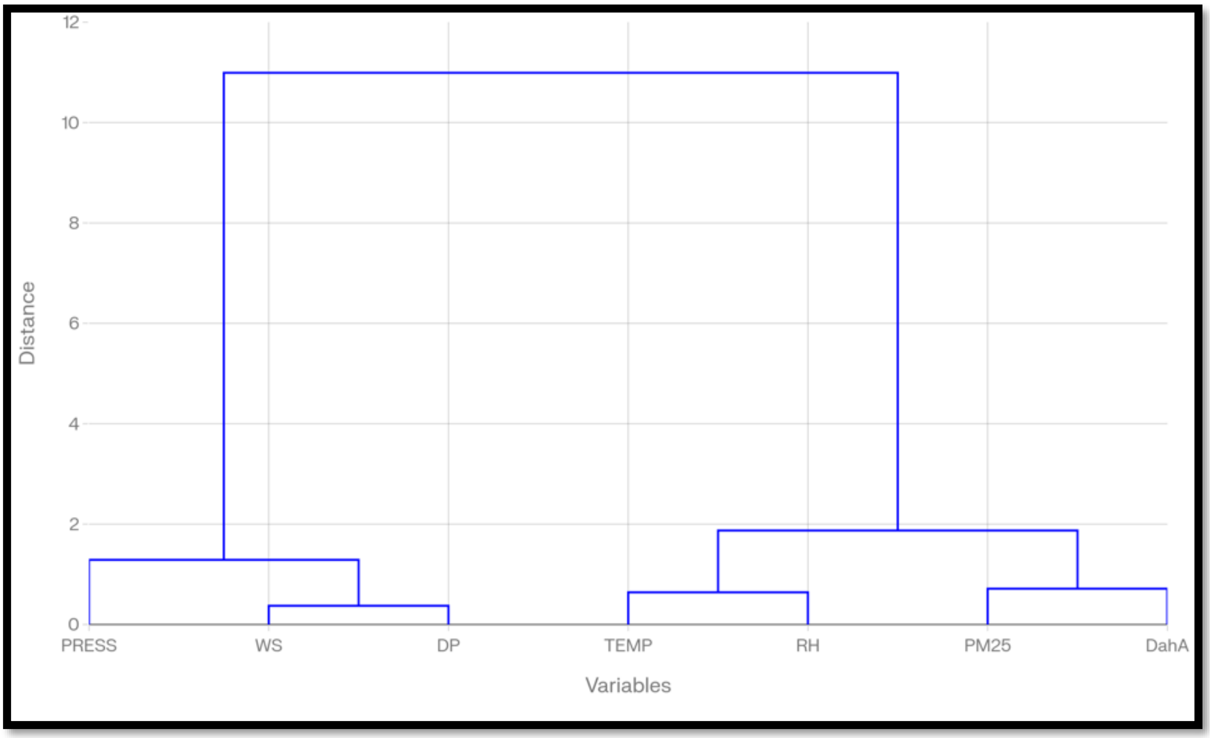


Figure 1. PCA biplot of PM2.5-bound dibenzo[a,h]anthracene and Microclimatic Parameters Dimensionality Reduction

In this study, two factors were identified using varimax with Kaiser Normalization from the PCA loadings. The two factors together explain 98.7% of the total variance in $PM_{2.5}$-bound DahA, with PC1 explaining 84.2% of the total variance and PC2 accounting for 14.5% respectively. PCA further resolved a dominant source–accumulation axis opposing a dispersion–removal axis, providing a systematic context for the ML-derived feature importance. The biplot presented in Figure 1 is in good agreement with the PCA. The clear separation of dry- and wet-season months along the primary principal component and the opposing loadings of source-related versus dispersion-related variables are consistent with biplot-based PCA interpretations of indoor–outdoor pollutant relationships in complex building environments [18].

.

© 2026 J. Chem. Allied Sci. | DOI: https://doi.org/10.60787/jcas.vol2no1.1| CC BY 4.0



HCA using Ward's linkage resolved two coherent, process-based groupings among $PM_{2.5}$-bound DahA and the microclimatic predictors (Figure 2). $PM_{2.5}$ clustered tightly with DahA, indicating a dominant source–carrier relationship in which fine particulate matter acts as the primary transport phase for high-molecular-weight PAHs in indoor environments, consistent with a source–accumulation axis. In contrast, microclimatic variables formed a distinct dispersion–removal grouping, with temperature and relative humidity reflecting thermodynamic and moisture controls on particle behaviour, and wind speed, dew point, and atmospheric pressure representing ventilation efficiency and indoor atmospheric stability [19]. This multivariate structure provides a systematic context for the machine learning–derived feature importance, corroborating that DahA variability is primarily governed by particulate loading, while microclimatic parameters modulate accumulation and removal processes. The dendrogram pattern is consistent with biplot-based PCA interpretations of pollutant–microclimate relationships in complex building environments [18]. These exploratory analyses provide contextual understanding but were not used directly to construct predictive models.

3.3 Machine Learning Development and Performance

Four different ML models – Decision Tree, Random Forest, Linear Regression and Generalized Linear Models were performed in Python (version 3.12) using standard scientific libraries from the data obtained in the study area [20,21].

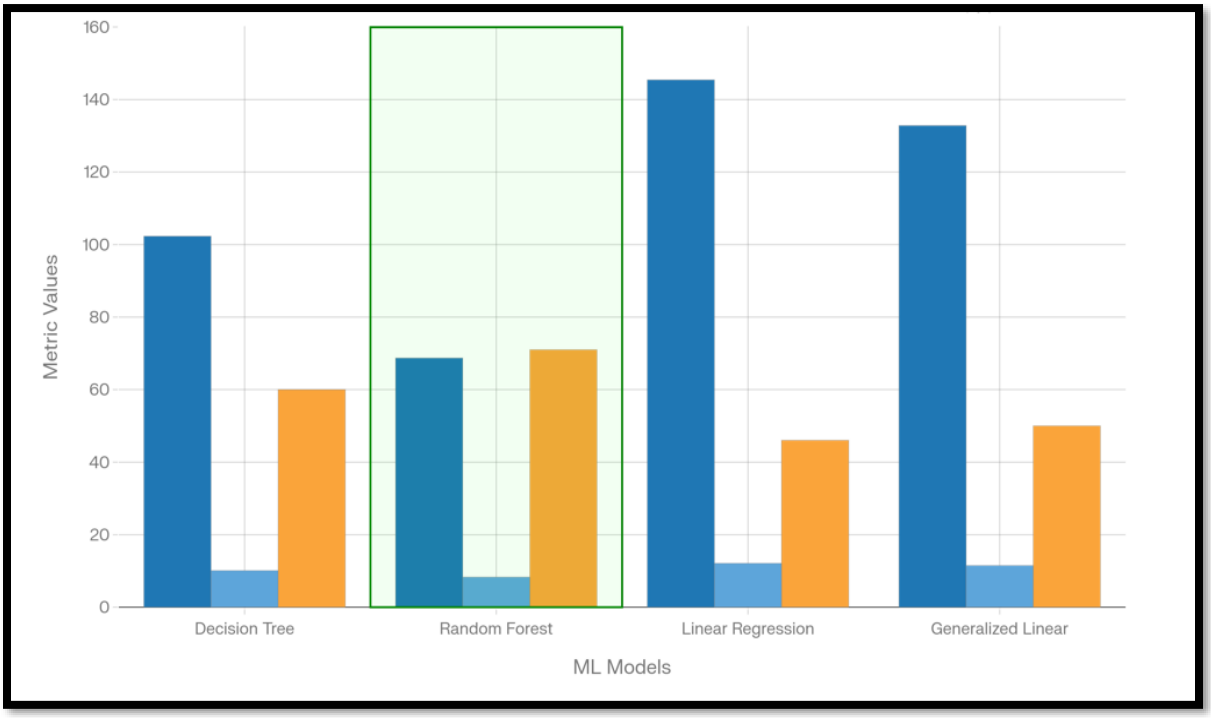


Figure 3. Model Performance Comparison for DahA

The model performance comparison (Figure 3) indicated that RF outperformed the other algorithms, achieving the highest predictive accuracy with an R^2 of 0.71, while DT, GLM, and LR had lower R^2 values of 0.60, 0.50, and 0.46, respectively. In addition to its superior accuracy, RF also exhibited the lowest root mean square error (RMSE), highlighting its robustness in predicting $PM_{2.5}$-bound DahA concentrations. Analysis of feature importance within the RF model revealed that $PM_{2.5}$ loading was the most influential predictor, followed in descending order by wind speed, relative humidity, atmospheric pressure, dew point, and temperature. This sequence highlights the primary role of particulate matter levels and meteorological factors in driving DahA concentration variability. This indicates that particulate mass is the dominant carrier of DahA, while ventilation proxies and moisture regulate accumulation and removal processes. These findings are in line with the previous ones, which prove the strength of tree-based ensembles in air-pollution modeling [8,9].

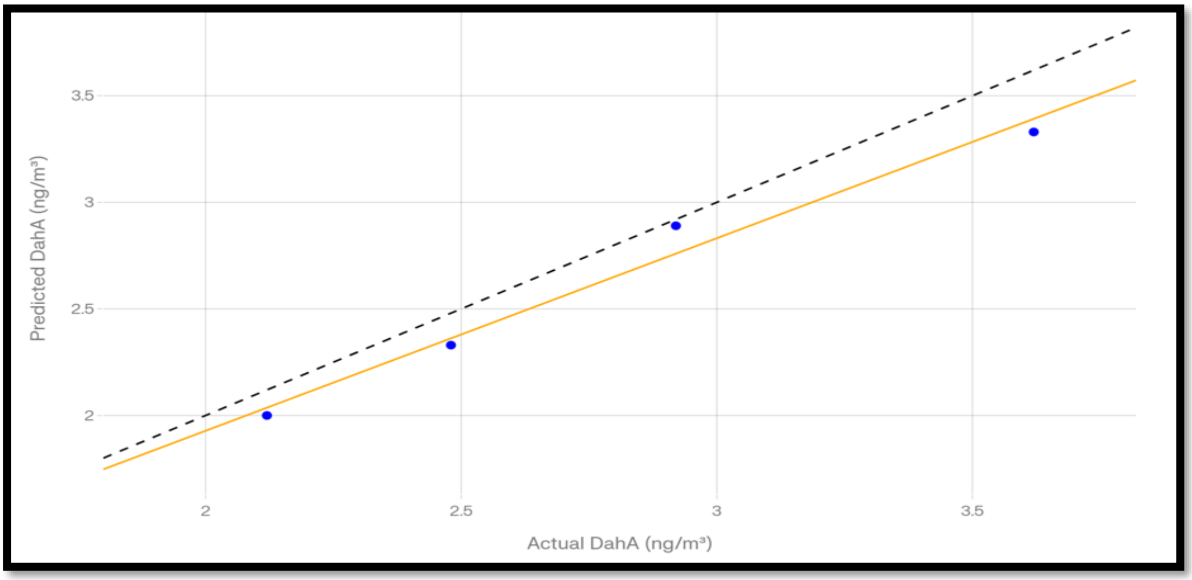


Figure 4. Random Forest- Actual Vs Predicted Values

The figure (4) shows the agreement between the concentrations of $PM_{2.5}$ bound to DahA measured by the empirical method and the machine-learning-based methods. The R^2 value for the RF model is 0.71, which indicates a strong correlation between the experimental and predicted data for DahA concentrations in the model. The higher R^2 value of 0.71 for RF suggests that it provides a better fit for predicting DahA concentrations in the study area. Random Forest methodology replicates the reported DahA variability with moderate-strong concordance, and has a better predictive accuracy [22] as indicated by higher coefficient of determination and closer association of predicted values to the 1:1 line of reference [23]. The above improved fit of the ensemble model is a result of the individual capacity of the model to induce nonlinear relationships between $PM_{2.5}$ loading and microclimatic predictors which are poorly represented in linear formulations. These results highlight the possibility of machine-learning models as decision support systems in real-time environmental management, either in the stabilization of soil or indoor air quality.

The machine-learning model created in this paper places more emphasis on the real-time microenvironmental operational control as opposed to seasonal classification hence shows that predictive performance is mostly based on the current conditions. Random Forest model proves to be robust in a heterogeneous indoor setting, which supports its applicability as an effective decision-support system to predict $PM_{2.5}$-bound DahA exposure in children places of worship, where polluted air areas are affected by the situation when combustion occurs and when the air is ventilated. The model can provide practical recommendations to reduce exposure, such as ventilation optimization, moisture control, and scheduling of activities, by classifying the most important drivers of microclimatic conditions, such as the loading of particles, moisture-controlling parameters, and ventilation proxies. This has been observed in relation to current research that demonstrates the wide applicability of machine-learning methods in predicting the environment. As an example, the framework of urban air-quality assessment and predictive mapping of health risks based on environmental, meteorological, and demographic data designed by Rajesh *et al.* [24] is a machine-learning-based model that works in real time. They have developed an approach utilizing a combination of the Random Forest and sophisticated algorithms, including Extreme Gradient Boosting (XGBoost) and Long Short-Term Memory network (LSTM), to provide high-temporal-resolution predictions and spatially stratified health-risks, thus demonstrating the usefulness of machine learning in dynamic exposure model. Equally, Zeini *et al.* [23] showed the effectiveness of the Random Forest models in capturing the nonlinear interactions in geopolymer-stabilized clay soils and also highlighted the ability of machine-learning to model non-deterministic heterogeneous environment systems where traditional deterministic methods are constrained.

The overall results of these studies re-establish that machine-learning models, especially the Random Forest model, can be very useful in the context of near-real-time and low costs of predicting environmental risk. In child-focused indoor environments, these strategies allow managing risks associated with air-quality through supplying quality forecasts of pollutant levels and informing mitigation strategies in environments with inadequate traditional monitoring facilities [25].

4.0 CONCLUSION

This study demonstrates that Random Forest regression ($R^2 = 0.71$) effectively predicts $PM_{2.5}$-bound dibenzo[a,h]anthracene concentrations in children's worship facilities using microclimatic variables. The findings indicate that particulate loading, ventilation proxies (wind speed), and moisture-related parameters (relative humidity and dew point) were identified as key drivers. The developed ML framework provides a transferable and cost-effective exposure assessment tool suitable for resource-limited environments. Targeted mitigation strategies, including ventilation optimization, moisture regulation, and controlled combustion activities, may significantly reduce indoor carcinogenic PAH exposure in child-focused environments.

Authors' Declarations

The authors affirm that the work presented is original and will accept all liability for any claims about the content.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Authors' Contributions

Archibong, U. D. and Michael, C. U. contributed to the literature search, data organization, and manuscript drafting. Both authors revised the manuscript for intellectual content, developed the conceptual framework, validated data, supervised the study, and coordinated the writing process. All authors approved the final version.

Acknowledgements: The authors acknowledge the use of Canva in designing the graphical abstract and Wiley's English Checker for English polishing. This research received no external funding at the time of submission.

References

[1] Choi, H., Harrison, R., Komulainen, H., Hites, R., Toriba, A., Hayakawa, K., et al. (2010). *Polycyclic aromatic hydrocarbons*. In WHO Guidelines for Indoor Air Quality: Selected Pollutants (pp. 61–120). Geneva: World Health Organization.

[2] Archibong, U. D., Okuo, J. M., & Agho, T. (2025). Characterization and modeling of indoor PM2.5-bound benzo[a]pyrene concentration in public schools: A comparative study of Oredo and Uhunmonde local government areas (LGAs), Edo State, Nigeria. *Pakistan Journal of Analytical & Environmental Chemistry, 26*(2), 240–256. https://doi.org/10.21743/pjaec/2025.12.06.

[3] Fang, G. C., Chang, K. F., Lu, C., & Bai, H. (2002). Toxic equivalency factors study of polycyclic aromatic hydrocarbons (PAHs) in Taichung City, Taiwan. *Toxicology and industrial health*, *18*(6), 279–288. https://doi.org/10.1191/0748233702th151oa.

[4] Mastral, A. M., Callén, M. S., García, T., & López, J. M. (2001). Benzo[a]pyrene, benzo[a]anthracene, and dibenzo[a,h]anthracene emissions from coal and waste tire energy generation at atmospheric fluidized bed combustion (AFBC). *Environmental Science & Technology, 35*(13), 2645–2649. https://doi.org/10.1021/es0015850.

[5] International Agency for Research on Cancer (IARC). (1987). *Some non-heterocyclic polycyclic aromatic hydrocarbons and related exposures* (Vol. 92, pp. 27–144). Lyon: IARC. [Monograph evaluating carcinogenic risks, includes dibenzo[a,h]anthracene (Group 2A)].

[6] Kanchana-at, T., Trivitayanurak, W., Chy, S., & Bordeerat, N. K. (2025). Particulate-Bound Polycyclic Aromatic Hydrocarbons and Heavy Metals in Indoor Air Collected from Religious Places for Human Health Risk Assessment. *Atmosphere*, *16*(6), 678. https://doi.org/10.3390/atmos16060678.

[7] Izevbigie, E. & Omagamre, W. (2026). Lung Cancer Risk Associated with $PM_{2.5}$-Bound Polycyclic Aromatic Hydrocarbons in a University Cafeteria in Southern Nigeria. 1-10. https://doi.org/10.21203/rs.3.rs-8507014/v1.

[8] Goudarzi, N., Shahsavani, D., Emadi-Gandaghi, F., & Arab Chamjangali, M. (2014). Application of random forests method to predict the retention indices of some polycyclic aromatic hydrocarbons. *Journal of Chromatography A, 1333*, 25–31. https://doi.org/10.1016/j.chroma.2014.01.048.

[9] Zhang, Y., Guo, Z., Peng, C., & Li, A. (2024). Random forest insights in prioritizing factors and risk areas of soil polycyclic aromatic hydrocarbons in an urban agglomeration area. *Science of the Total Environment, 957*, 177583. https://doi.org/10.1016/j.scitotenv.2024.177583.

[10] Salleh, S. F., Suleiman, A. A., Daud, H., Othman, M., Sokkalingam, R., & Wagner, K. (2023). Tropically Adapted Passive Building: A Descriptive-Analytical Approach Using Multiple Linear Regression and Probability Models to Predict Indoor Temperature. *Sustainability*, *15*(18), 13647. https://doi.org/10.3390/su151813647.

[11] World Health Organization. (2021). WHO global air quality guidelines: Particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulphur dioxide and carbon monoxide. World Health Organization.

[12] United States Environmental Protection Agency. (2022). NAAQS Table. Last updated February 7, 2024. [Accessed 2024].

[13] Archibong, U. D. & Okuo, J. M. (2024). Correlation of Fine Particulates and Meteorological Parameters in Indoor Public Schools Environment within Benin City, Nigeria. *Journal of Materials & Environmental Sustainability Research, 4*(2), 30–38. https://doi.org/10.55455/jmesr.2024.006.

[14] Guo, L. C., Bao, L. J., She, J., & Zeng, E. Y. (2014). Significance of wet deposition to removal of atmospheric particulate matter and polycyclic aromatic hydrocarbons: A case study in Guangzhou, China. *Atmospheric Environment, 83*, 136–144. https://doi.org/10.1016/j.atmosenv.2013.11.012.

[15] Taylor, E. T., Wirmvem, M. J., Sawyerr, V. H., & Nakai, S. (2015). Characterization and determination of PM2.5-bound polycyclic aromatic hydrocarbons (PAHs) in indoor and outdoor air in Western Sierra

Leone. *Journal of Environmental & Analytical Toxicology,* 5(5), Article 307. https://doi.org/10.4172/2161-0525.1000307.

[16] Wang, D., Wu, S., Gong, X., Ding, T., Lei, Y., Sun, J., & Shen, Z. (2023). Characterization and Risk Assessment of $PM_{2.5}$-Bound Polycyclic Aromatic Hydrocarbons and their Derivatives Emitted from a Typical Pesticide Factory in China. *Toxics*, *11*(7), 637. https://doi.org/10.3390/toxics11070637.

[17] Wang, M., Jia, S., Lee, S. H., Chow, A., & Fang, M. (2021). Polycyclic aromatic hydrocarbons (PAHs) in indoor environments are still imposing carcinogenic risk. *Journal of Hazardous Materials, 409*, 124531. https://doi.org/10.1016/j.jhazmat.2020.124531.

[18] Zhang, H., & Srinivasan, R. (2021). A biplot-based PCA approach to study the relations between indoor and outdoor air pollutants using case study buildings. *Buildings,* *11*(5), 218. https://doi.org/10.3390/buildings11050218.

[19] Gao, X., Wang, Z., Sun, X., Gao, W., Jiang, W., Wang, X., Zhang, F., Wang, X., Yang, L., & Zhou, Y. (2024). Characteristics, source apportionment and health risks of indoor and outdoor fine particle-bound polycyclic aromatic hydrocarbons in Jinan, North China. *PeerJ*, *12*, e18553. https://doi.org/10.7717/peerj.18553.

[20] Hui, X., Guo, S., Shi, X., Yang, W., Pan, J., & Gao, H. (2024). Machine learning-based analysis and prediction of meteorological factors and urban heatstroke diseases. *Frontiers in Public Health, 12*, 1420608. https://doi.org/10.3389/fpubh.2024.1420608.

[21] Racić, N., Ružičić, S., Petrić, V., Terzić, T., Antunović, M., Škaro, I., et al. (2026). Assessment of contributors to airborne PAHs and heavy metals in PM_{10} using temporal, spatial, traffic and heating data in explainable machine learning models. *Atmospheric Environment: X, 29*, 100413. https://doi.org/10.1016/j.aeaoa.2026.100413.

[22] Yenkikar, A., Mishra, V. P., Bali, M., & Ara, T. (2025). Explainable forecasting of air quality index using a hybrid random forest and ARIMA model. *MethodsX, 15*, 103517. https://doi.org/10.1016/j.mex.2025.103517.

[23] Zeini, H. A., Al-Jeznawi, D., Imran, H., Bernardo, L. F. A., Al-Khafaji, Z., & Ostrowski, K. A. (2023). Random Forest Algorithm for the Strength Prediction of Geopolymer Stabilized Clayey Soil. Sustainability, 15(2), 1408. https://doi.org/10.3390/su15021408.

[24] Rajesh, M., Babu, R. G., Moorthy, U., & Easwaramoorthy, S. V. (2025). Machine learningdriven framework for realtime air quality assessment and predictive environmental health risk mapping. Scientific reports, 15(1), 28801. https://doi.org/10.1038/s41598-025-14214-6.

[25] Yang, X., Li, Y., Liu, L., & Zang, Z. (2025). Prediction of respiratory diseases based on random forest model. *Frontiers in public health*, *13*, 1537238. https://doi.org/10.3389/fpubh.2025.1537238.