KARYA ILHAM PUBLISHING
CONNECTING MINDS, BRIDGING IDEAS, INSPIRING READERS

Journal of Advanced Research in Computing and Applications

Journal homepage:
https://karyailham.com.my/index.php/arca/index
ISSN: 2462-1927

Air quality index classification based on machine learning techniques: A case study from Hangzhou City

Zalinda Othman[1,*], Yue Li[1]

[1] Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 45600 UKM Bangi, Selangor, Malaysia

ARTICLE INFO

Article history:
Received 2 October 2026
Received in revised form 28 August 2026
Accepted 31 August 2026
Available online 3 September 2026

Keywords:
Multi-class classification; Air Quality Index (AQI); Machine Learning; LSTM

ABSTRACT

Atmospheric pollutants pose a significant threat to human health and ecological stability, necessitating the development of accurate and reliable classification models to support informed decision-making. This study leverages state-of-the-art machine learning techniques to classify air quality levels in Hangzhou. Through comprehensive data analysis, it examines the impact of data preprocessing and feature engineering on classification performance. It evaluates the effectiveness of five machine learning models: Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), CatBoost, and Long Short-Term Memory (LSTM) networks. To enhance data integrity, preprocessing techniques such as anomaly detection, missing value imputation, and the Synthetic Minority Oversampling Technique (SMOTE) were applied. At the same time, feature engineering strategies, including seasonal and temporal feature extraction, were implemented to improve classification accuracy. Experimental results show that Random Forest achieved the highest classification accuracy and PR-AUC, surpassing XGBoost and CatBoost, demonstrating strong performance. While LSTM benefited from seasonal feature integration, its classification effectiveness remained lower than that of ensemble models, particularly in handling class imbalances. SVM exhibited moderate performance but struggled with complex data distributions. These findings highlight the strengths and limitations of each model, offering a scalable classification framework that can support urban air quality management and be adapted to similar environments.

1. Introduction

Air pollution remains a significant global environmental and public health challenge, causing detrimental effects on human health, ecosystems, and climate stability [1]. Rapid urbanization and industrialization have exacerbated atmospheric pollutant concentrations, increasing the incidence of respiratory illnesses, cardiovascular diseases, and environmental degradation [2,3]. Traditional air quality classification methods, which primarily rely on empirical and statistical models, are often limited in capturing complex and nonlinear pollutant interactions, thereby hindering effective

* *Corresponding author.*
E-mail address: zalinda@ukm.edu.my

environmental management and decision-making processes [4]. Machine learning (ML) has emerged as an effective alternative due to its robust ability to analyze complex, multidimensional environmental data. Several ML techniques, including ensemble methods and deep learning models, have been successfully applied to environmental classification tasks, demonstrating superior performance in prediction accuracy and interpretability [5]. However, challenges remain in addressing data imbalance, anomalies, and integrating relevant temporal and seasonal features, significantly influencing classification performance [6]. Motivated by these considerations, this study evaluates the effectiveness of five advanced models: Random Forest (RF), Extreme Gradient Boosting (XGBoost), CatBoost, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM), for classifying air quality levels in Hangzhou, China. This research investigates explicitly data preprocessing approaches such as anomaly detection, missing value imputation, and the Synthetic Minority Oversampling Technique (SMOTE) to address data imbalance. Additionally, seasonal and temporal feature engineering strategies were integrated to enhance classification accuracy further. The findings from this study aim to provide practical insights and robust modeling frameworks that support urban air quality management and policymaking.

2. Methodology

2.1 Study Area

Hangzhou, the capital city of Zhejiang Province, is in the eastern region of China and is renowned for its rapid economic growth, technological innovation, and urban expansion. The rapid development of Hangzhou has led to increased environmental pressures, particularly on air quality. The primary sources of atmospheric pollutants in Hangzhou include industrial emissions, vehicular exhaust, construction activities, and regional transport of pollutants [7]. These factors have contributed to fluctuating air quality, presenting significant management challenges. Furthermore, the city's prominence as a host of international events, such as the G20 Summit in 2016 and the Asian Games in 2022, underscores the importance of effective air quality monitoring and management systems. This context makes Hangzhou an ideal study area for assessing the effectiveness and applicability of advanced ML-based air quality classification models.

2.2 Data Collection

Air quality data for 2023, recorded hourly, was acquired from environmental monitoring stations in Hangzhou. The dataset comprises pollutant measurements for particulate matter (PM2.5, PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), ozone (O3), and carbon monoxide (CO), key indicators affecting urban air quality (Zhang et al., 2024). The dataset spans January to December 2023, offering a robust temporal resolution for analyzing pollutant fluctuations and trends throughout different seasons. Table 1 summarizes the statistical description of pollutant concentrations measured during the study period, providing essential insights into the variability and distribution characteristics of Hangzhou's air quality dataset.

Table 1
Air quality parameters and their analysis

Parameter	Description	Unit
AQI (Air Quality Index)	To measure how clean or polluted the air is.	-
PM2.5	Tiny particles (< 2.5 microns) are mostly from industry and traffic. Long-term exposure increases the risk of lung and heart problems.	μg/m3
PM2.5_24h	24-hour average of PM2.5, a key air pollution indicator linked to health.	μg/m3
PM10	Tiny particles (< 10 microns) from construction and road dust accumulate in low wind speeds. It should be analyzed with meteorological factors.	μg/m3
PM10_24h	24-hour average of PM10, used for assessing long-term air pollution and health impact.	μg/m3
SO2 (Sulfur Dioxide)	Emitted mainly from coal combustion and industry. Short-term exposure causes respiratory irritation; long-term exposure harms ecosystems.	μg/m3
SO2_24h	The 24-hour average of sulfur dioxide is essential for assessing industrial pollution and wind influence.	μg/m3
NO2 (Nitrogen Dioxide)	Mainly from vehicles and industrial emissions. High levels are found in traffic-dense areas; best analyzed with traffic flow data.	μg/m3
NO2_24h	24-hour average of nitrogen dioxide, helpful in analyzing traffic-related pollution hotspots.	μg/m3
O3 (Ozone)	A secondary pollutant formed by sunlight reactions with NO2 and VOCs. High levels cause respiratory issues and environmental damage.	μg/m3
O3_24h	24-hour average of ozone, helpful in assessing prolonged exposure risks, especially in urban areas.	μg/m3
O3_8h	8-hour average of ozone, used as a key indicator for short-term exposure and its respiratory health impact. Often used in air quality regulations.	μg/m3
CO (Carbon Monoxide)	A colorless, odorless gas produced by burning fossil fuels. High levels can cause poisoning and respiratory issues.	mg/m3
CO_24h	The 24-hour average of carbon monoxide is important for assessing long-term exposure risks.	mg/m3

An initial statistical analysis was conducted to comprehensively characterize pollutant distributions during the study period, focusing on essential metrics such as mean, standard deviation, minimum, maximum, and quartile ranges. As presented in Table 2, the mean concentrations indicate that PM_{10} (54.35 μg/m3) and $PM_{2.5}$ (31.21 μg/m3) were the predominant pollutants, reflecting the general conditions of particulate pollution in Hangzhou until 2023. The relatively high standard deviations for these pollutants (PM_{10}: 39.24 μg/m3; $PM_{2.5}$: 21.18 μg/m3) reveal considerable variability, suggesting the presence of episodic pollution events or significant seasonal fluctuations. The range of minimum and maximum values further underscores the occurrence of extreme pollution events. For instance, PM_{10} levels ranged from a minimum of 4 μg/m3 to 589 μg/m3, suggesting potential links to typical industrial activities or unfavorable meteorological conditions, such as temperature inversions. The application of quartile analysis yielded further evidence of skewed pollutant distributions, underscoring the necessity for normalization and logarithmic transformation during subsequent preprocessing stages. The amalgamation of these detailed statistical insights has informed the selection and implementation of rigorous data preprocessing

techniques, including anomaly detection, missing value imputation, and data normalization, thereby enhancing the robustness and reliability of the modeling outcomes.

Table 2
Summary statistics of air quality parameters

Parameter	Count	Mean	Std Dev	Min	25%	50%	75%	Max
AQI	8710	52.16	30.83	9	32	48	63	489
PM2.5	8715	31.21	21.18	2	17	27	39	173
PM2.5_24h	8723	31.16	18.85	5	19	28	38	158
PM10	8714	54.35	39.24	4	30	46	67	589
PM10_24h	8721	54.24	34.88	7	32	48	67	408
SO2	8728	6.47	1.23	4	6	6	7	19
SO2_24h	8726	6.45	1.01	5	6	6	7	11
NO2	8721	30.07	17.64	3	17	25	38	120
NO2_24h	8722	30.14	14.87	7	19	26	39	94
O3	8731	63.29	44.76	2	29	55	85	238
O3_24h	7474	127.91	48.54	15	90	120	163	245
O3_8h	8675	60.52	41.05	0	30.5	55	85	214
CO	8723	0.65	0.15	0.31	0.54	0.63	0.73	1.61
CO_24h	8721	0.65	0.13	0.31	0.56	0.63	0.72	1.38

2.3 Data Processing

Comprehensive data processing was crucial for ensuring the quality and reliability of the air quality classification models. To enhance dataset consistency and model performance, this study systematically addressed data-related issues, including missing values, anomalies, distribution skewness, and class imbalances. Initially, missing values frequently arising from sensor malfunctions or routine maintenance interruptions were imputed using polynomial regression interpolation (second order) to preserve the nonlinear dynamics of pollutant data. For features with pronounced daily cyclical patterns, such as AQI, missing observations were imputed with a seven-day rolling average, maintaining temporal continuity and capturing short-term trends [6]. Outlier detection and correction were subsequently implemented to minimize the distortion effects from anomalous values, which could originate from equipment failures or episodic pollution events. The Z-score method ($|Z| > 3$) effectively identified extreme pollutant concentrations, with median replacement applied to pollutants exhibiting fewer anomalies, such as PM2.5 and SO2. In contrast, linear interpolation was applied to features experiencing frequent anomalies, including O3 and CO, preserving dataset coherence without introducing bias [8]. Data normalization was done through Min-Max scaling to standardize pollutant concentrations within a consistent range [0,1]. Additionally, highly skewed pollutant features, particularly PM2.5, PM10, and NO2, underwent logarithmic transformation, stabilizing variance and reducing the negative impact of extreme values on model accuracy. This transformation significantly reduced feature skewness, improving data symmetry and enhancing subsequent predictive modeling [9].

The target variable, representing air quality categories such as "Good", "Moderate" and "Hazardous", was numerically encoded to facilitate compatibility with machine learning algorithms. Each AQI category was mapped to a numeric label as shown in Table 3: "Good" as 0, "Moderate" as 1, "Unhealthy for Sensitive Groups" as 2, "Unhealthy" as 3, "Very Unhealthy" as 4, and "Hazardous" as 5. This encoding ensured consistency and enabled efficient model training and evaluation while supporting straightforward interpretation and reversibility of model outcomes.

Table 3
Numerical encoding of air quality index (AQI) Categories

Air Quality Category	Encoded Value
Good	0
Moderate	1
Unhealthy for Sensitive Groups	2
Unhealthy	3
Very Unhealthy	4
Hazardous	5

Class imbalance presented a notable challenge due to the disproportionate distribution of air quality categories within the dataset, particularly underrepresenting higher pollution classes like "Very Unhealthy" and "Hazardous." To counteract this imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied during model training to generate synthetic examples of minority classes, thereby providing balanced learning opportunities for all categories. This method improved models' capacity to predict rare but critical pollution events [10] accurately. To verify the effectiveness of the data processing procedures, visual analytics methods including boxplots and histograms were utilized [6,11]. These visual assessments confirmed a significant reduction in data skewness, anomalies, and missing values post-processing, ensuring enhanced data integrity for subsequent modeling tasks [12]. Furthermore, stratified sampling was employed during dataset splitting into training and testing subsets to preserve original class proportions, facilitating unbiased model evaluation and robust generalization assessments [13].

2.4 Feature Engineering

Effective feature engineering significantly enhances the predictive capabilities of machine learning models by selecting and transforming critical variables to represent air pollution dynamics [14] optimally. In this study, correlation analysis was conducted to identify the most relevant pollutant indicators associated with air quality categories. A correlation threshold of $|r| \geq 0.5$ was utilized to retain informative variables, ensuring that only highly significant predictors were included, reducing redundancy and improving model interpretability [15]. The correlation analysis shown in Figure 1 highlighted strong associations between AQI categories and several key pollutants, specifically PM2.5, PM10, NO2, and their respective 24-hour moving averages (PM2.5_24h, PM10_24h, NO2_24h), all of which demonstrated correlation coefficients ranging from 0.52 to 0.76. These pollutants are well-documented for their substantial contributions to air quality deterioration and public health risks [16,17].

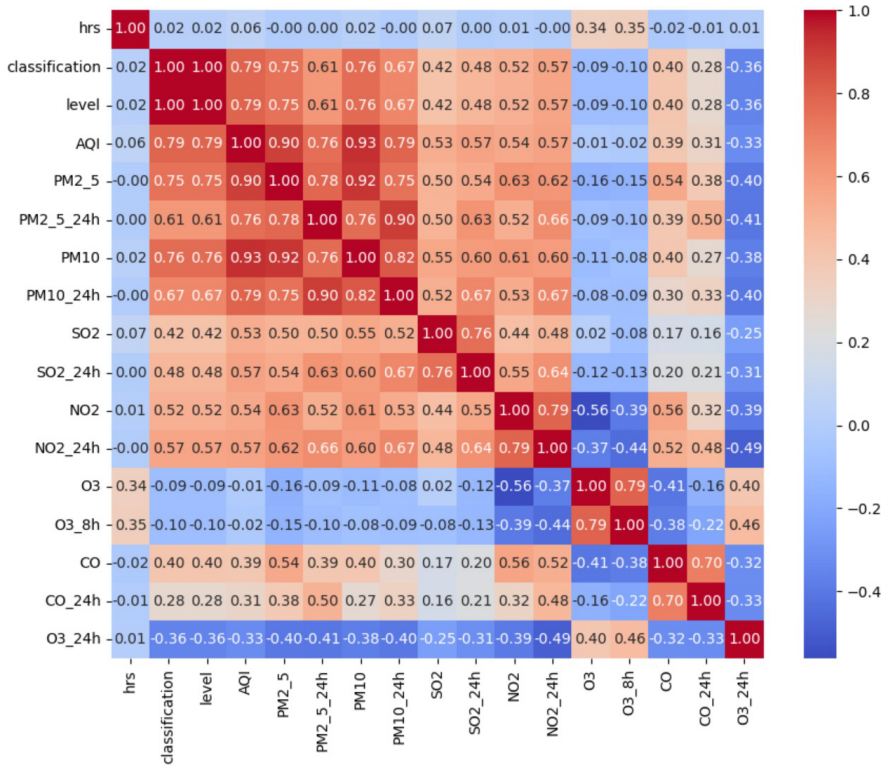


Fig. 1. Feature correlation heatmap

Additional temporal features were derived from the timestamp dataset to capture temporal variations critical to air pollution modeling. The hourly timestamps were systematically transformed into temporal attributes, including categorical representations of seasons (spring, summer, autumn, winter), thus enabling models to identify cyclical and seasonal pollution patterns better. Previous research has highlighted the importance of incorporating temporal features into air quality prediction models, especially when meteorological data are unavailable [6]. Furthermore, the dataset was structured using a sliding window technique to enhance the sequential modeling capability for temporal models such as LSTM, creating input-output sequences consisting of three consecutive time steps. This data structuring allowed models to capture short-term temporal dependencies and fluctuations, improving predictive accuracy for air quality classification tasks [18]. To prevent data leakage and maintain robust model generalization, features directly representing the target outcomes, such as AQI numerical values and categorical classification labels, were excluded from the feature set [14,17]. This step ensured unbiased model training and evaluation while preserving predictive integrity. The final feature set, combining critical pollutant concentrations, their moving averages, and engineered temporal attributes, provided a robust basis for accurate classification and interpretability across different machine learning models [19].

2.5 Machine Learning Model

This study employed five advanced machine learning techniques: Random Forest (RF), Extreme Gradient Boosting (XGBoost), CatBoost, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) to classify air quality levels. Each model was selected based on its distinctive strengths in addressing complex, nonlinear patterns and imbalanced datasets common in environmental classification problems. RF is an ensemble method widely used due to its ability to model complex interactions among features, high robustness to overfitting, and provision of interpretable feature importance measures [20]. It integrates the predictions of numerous decision

trees constructed through bootstrap sampling, significantly enhancing classification accuracy. XGBoost is another ensemble learning model, recognized for its efficiency in nonlinear relationships and robustness against imbalanced and sparse data distributions through iterative decision tree boosting. Regularization techniques embedded in XGBoost effectively mitigate model complexity, providing accurate predictions for challenging environmental datasets [21]. CatBoost is a gradient boosting algorithm specifically designed to handle categorical data efficiently. Its unique ordered target-based encoding method effectively reduces target leakage and overfitting, making CatBoost suitable for structured and imbalanced datasets common in air quality studies [22]. SVM was selected as a baseline classifier to provide comparative insights into linear and nonlinear decision boundaries through kernel transformations. Despite its robustness and general applicability in numerous classification tasks, SVM struggles with highly imbalanced and complex nonlinear distributions, posing challenges in accurate air quality categorization [23]. LSTM networks, a deep-learning approach, exploited temporal dependencies and sequential patterns inherent within the hourly air quality data. Given their capacity to learn and memorize temporal fluctuations, LSTM models provided valuable insights into the time-dependent nature of pollutant variations and were enhanced by incorporating engineered temporal features [5,24]. Hyperparameters for all models were rigorously optimized through grid search and Bayesian optimization methods combined with stratified 5-fold cross-validation. Key hyperparameters optimized included the number and depth of trees (RF, XGBoost, CatBoost), kernel type and regularization parameters (SVM), and number of units, dropout rates, and learning rates (LSTM).

2.6 Model Evaluation

Multiple evaluation metrics were employed to assess model performance in classifying air quality levels, considering the imbalanced nature of the dataset. Accuracy alone was insufficient, as it tends to overestimate performance in datasets with a dominant majority classes. Instead, Precision-Recall Area Under Curve (PR-AUC) was selected as the primary metric, providing a more reliable assessment of classification effectiveness, particularly for minority classes [25]. The PR-AUC score, Macro PR-AUC and Micro PR-AUC are computed as follows:

$$PR - AUC = \int_0^1 P(R)dR \quad (1)$$

$$Macro\ PR - AUC = \frac{1}{n}\sum_{k=1}^{n} PR - AUC_k \quad (2)$$

$$Micro\ PR - AUC = PR - AUC_{global} \quad (3)$$

where P(R) represents precision as a function of recall. Macro PR-AUC and Micro PR-AUC were calculated to evaluate model performance across different air quality categories, ensuring a balanced comparison between models. Macro PR-AUC equally weights each class, capturing performance on minority pollution levels, while Micro PR-AUC aggregates all instances, favoring overall classification effectiveness.

Complementary metrics, including F1-score, Precision, Recall, and Accuracy, were computed to analyze model behavior further:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)} \quad (5)$$

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)} \quad (6)$$

$$Accuracy = \frac{Number\ of\ Correct\ Classifications}{Total\ Classifications} \quad (7)$$

The dataset was split into 70% training and 30% testing subsets to ensure fair model evaluation, maintaining class distribution via stratified sampling. For the LSTM model, which relies on sequential dependencies, a chronological train-test split was used to prevent information leakage. 5-fold cross-validation was applied to RF, XGBoost, CatBoost, and SVM models to enhance evaluation robustness further, minimizing variance in model performance estimation. Additionally, hyperparameter tuning was conducted using Bayesian optimization and grid search, optimizing key parameters such as tree depth (RF, XGBoost, CatBoost), kernel functions (SVM), and learning rate/dropout rate (LSTM). Confusion matrices supported performance comparison, offering detailed insights into misclassification patterns. This evaluation framework, underpinned by a well-defined workflow including data cleaning, correlation-based feature selection, temporal feature engineering, stratified train-test splitting, hyperparameter optimization, and multi-metric model evaluation, ensured a comprehensive, unbiased, and reproducible assessment of model effectiveness in air quality classification [26,27].

3. Results

This section discusses the results obtained from the surface pressure measurement The classification results demonstrate the effectiveness of different machine learning models in predicting air quality levels based on pollutant concentrations and engineered features. The optimization models were evaluated using a comprehensive performance framework. Macro PR-AUC and Micro PR-AUC were emphasized as primary metrics due to their ability to assess model performance across imbalanced classes. As shown in Table 4, RF emerged as the best-performing model, achieving the highest Macro PR-AUC of 0.9839 and Micro PR-AUC of 0.9962. These metrics highlight RF's exceptional ability to balance precision and recall across all air quality categories while minimizing misclassifications at the overall dataset level. This superior performance is attributed to RF's ensemble structure, which effectively captures complex, nonlinear interactions and handles high-dimensional feature spaces [9]. Additionally, RF's strong F1-score of 0.9727 further underscores its robustness in multi-class classification tasks.

Table 4
Performance comparison of machine learning models

Model	Macro PR-AUC	Micro PR-AUC	F1-Score	Precision	Recall	Accuracy
RF	0.9839	0.9962	0.9727	0.9735	0.9725	0.9725
XGBoost	0.9689	0.9962	0.9743	0.9748	0.9741	0.9741
CatBoost	0.9651	0.9960	0.9735	0.9740	0.9733	0.9733
LSTM	0.8903	0.9787	0.9350	0.9365	0.9344	0.9344
SVM	0.8404	0.9442	0.8666	0.8732	0.8638	0.8638

Closely following RF, both XGBoost and CatBoost demonstrated competitive results. XGBoost recorded a Macro PR-AUC of 0.9689 and a Micro PR-AUC of 0.9962, while CatBoost achieved a Macro PR-AUC of 0.9651 and a Micro PR-AUC of 0.9960. These results affirm the strong generalization capabilities of gradient boosting-based models, which optimize classification performance through iterative learning. Their respective F1-scores of 0.9743 (XGBoost) and 0.9735 (CatBoost) indicate a consistent ability to handle complex variable relationships and maintain balanced classification, particularly in class imbalance. The LSTM model, while not outperforming the ensemble models,

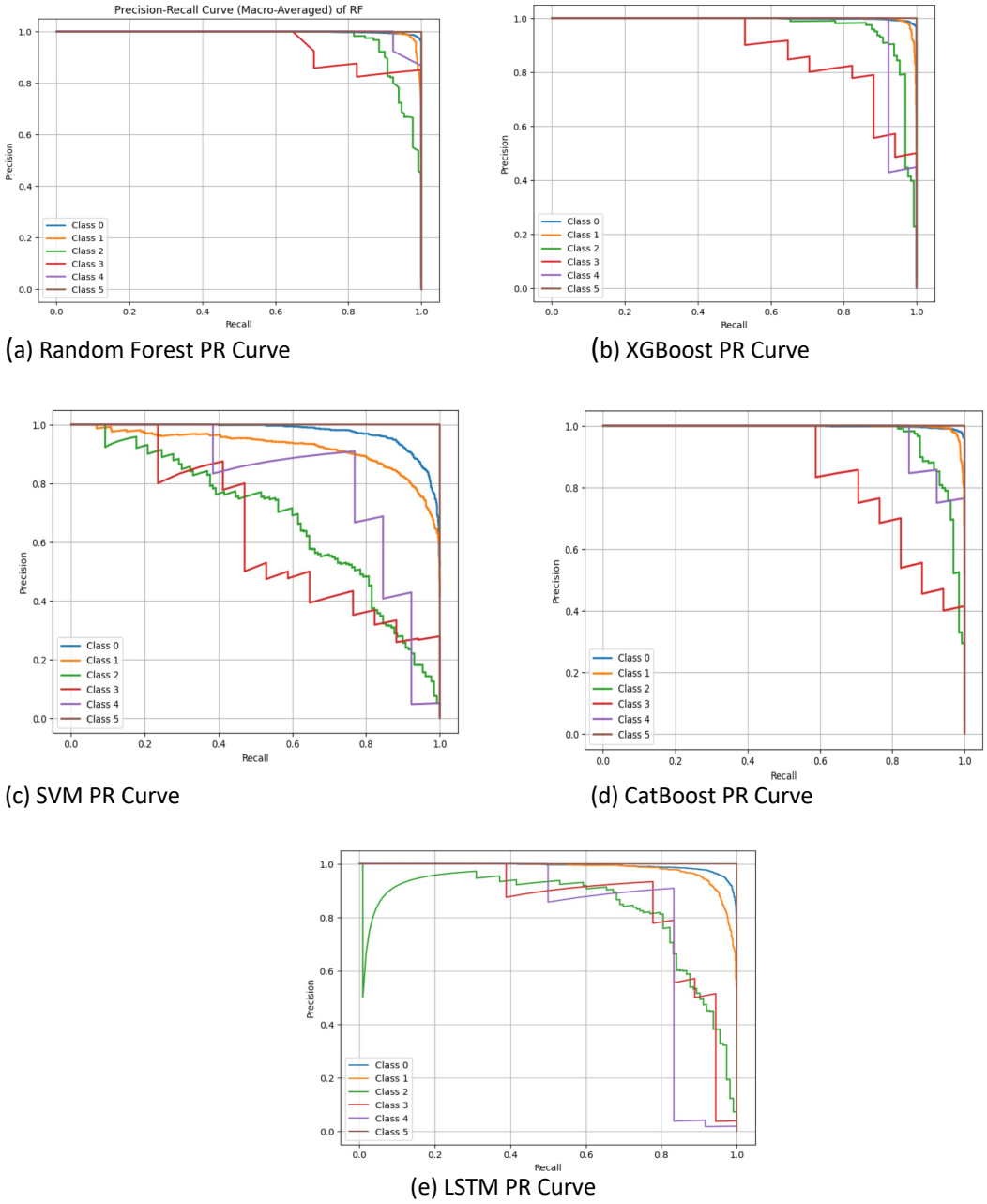
demonstrated unique strengths in modeling sequential patterns and temporal dependencies. Its Macro PR-AUC of 0.8903 and Micro PR-AUC of 0.9787 reflect reasonable classification performance, mainly when supported by engineered temporal features designed to capture cyclical trends in air quality [28]. Conversely, the Support Vector Machine (SVM) model exhibited the weakest performance among the five models. With a Macro PR-AUC of 0.8404 and a Micro PR-AUC of 0.9442, SVM struggled to maintain classification balance, particularly for minority AQI categories. Its F1-score of 0.8666 confirms its limited capability to capture complex feature interactions and adapt to class imbalance. It is consistent with previous findings that SVMs require advanced kernel tuning and feature engineering to improve classification accuracy in such scenarios [29].

The comparison between LSTM models with and without seasonal features, as shown in Table 5, reveals a modest improvement in key metrics, suggesting that temporal feature engineering can enhance performance in sequence-based models. Nevertheless, LSTM's F1-score of 0.9350 was significantly lower than that of ensemble models, indicating its relative limitations in handling highly imbalanced datasets without additional temporal or meteorological inputs.

Table 5
Comparison of the LSTM model with and without season features

Metric	LSTM	LSTM (With Season)
Accuracy	0.9344	0.9351
Macro PR-AUC	0.8903	0.8993
Micro PR-AUC	0.9787	0.9772
Macro Avg F1	0.8624	0.8643
Weighted Avg F1	0.9350	0.9355

In addition to the quantitative metrics provided in Table 4, Precision-Recall (PR) curves were analyzed to offer deeper insight into each model's performance across different thresholds, especially in class imbalance. While metrics such as Macro PR-AUC and F1 score summarize overall performance, PR curves allow a more nuanced visualization of how well models balance precision and recall for majority and minority AQI categories at various decision boundaries. As illustrated in Figure 2, RF exhibits a near-perfect PR curve, maintaining consistently high precision and recall values across all thresholds. This visual trend confirms its superior robustness in identifying minority classes without sacrificing accuracy on dominant categories. Similarly, both XGBoost and CatBoost show smooth, stable PR curves that closely follow RF's performance line, demonstrating their ability to effectively capture complex feature interactions while mitigating the risks of overfitting to majority classes. These smooth gradients suggest ensemble models are well-suited for multi-class classification tasks with imbalanced data distributions, aligning with their iterative learning and ensemble averaging mechanisms.



(a) Random Forest PR Curve

(b) XGBoost PR Curve

(c) SVM PR Curve

(d) CatBoost PR Curve

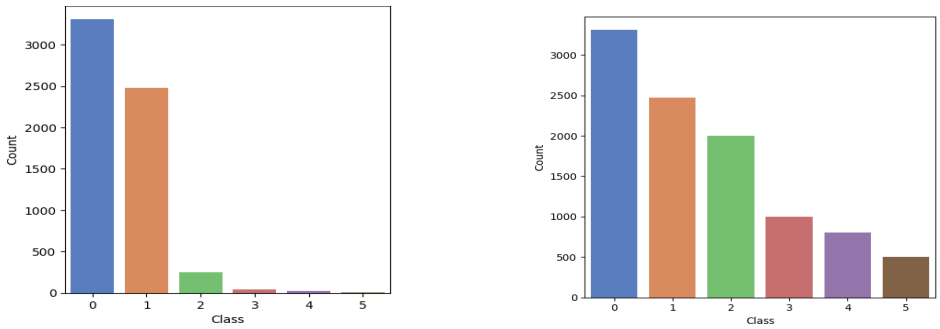
(e) LSTM PR Curve

Fig. 2. Precision-Recall curves for all evaluated models: (a) Random Forest; (b) XGBoost; (c) SVM; (d) CatBoost; (e) LSTM

In comparison, the LSTM model's PR curve displays moderate fluctuations. Although it maintains reasonably high precision and recall in middle thresholds, a slight dip at more extreme cutoffs indicates challenges in confidently classifying edge cases, particularly for the rarest AQI categories. This observation aligns with LSTM's temporal modeling advantage and highlights its relative sensitivity to noise and sequence length constraints. The SVM model shows the most pronounced decline in its PR curve, with noticeable drops in precision at lower thresholds. The result suggests that SVM struggles to maintain stable classification performance when dealing with underrepresented categories and is prone to higher false positive rates in borderline scenarios. The steeper slope of the SVM curve not only confirms its numerical inferiority seen in Table 4 and visually emphasizes its limitations in capturing nonlinear interactions and handling class imbalance without additional kernel or feature adjustments. Further analysis of confusion matrices indicated that most

misclassifications occurred between adjacent air quality categories, particularly between "Moderate" and "Unhealthy" classes. This phenomenon likely results from gradual pollutant concentration changes and measurement uncertainties, which blur classification boundaries [6]. These patterns emphasize the potential value of refining feature selection strategies and hyperparameter tuning to improve model performance in borderline classification cases. In conclusion, Random Forest demonstrated the highest overall effectiveness, balancing Macro PR-AUC, Micro PR-AUC, and F1-score with outstanding consistency. XGBoost and CatBoost also proved highly effective, reinforcing the superiority of ensemble tree-based models for complex air quality classification tasks. Although LSTM offered advantages in modeling sequential dependencies, its overall classification performance remained moderate compared to ensemble methods [30]. SVM, constrained by its inability to handle imbalanced and nonlinear data structures, was the least suitable model without substantial optimization. These findings highlight the importance of selecting robust ensemble models and complementing them with appropriate feature engineering and imbalance handling techniques to achieve accurate and reliable air quality classification [8,9].

Given the imbalanced nature of the dataset, class imbalance handling played a crucial role in model performance. Synthetic Minority Oversampling Technique (SMOTE) was applied to mitigate the effects of underrepresented AQI categories, particularly "Very Unhealthy" and "Hazardous", which comprised a small fraction of the dataset. Models trained with SMOTE exhibited improved recall for minority classes, as reflected in Figure 3, where the distribution of class samples shows noticeable changes before and after applying SMOTE. The minority classes, which were previously underrepresented, were effectively augmented, resulting in a more balanced dataset. This improvement reduces the risk of model bias towards majority classes and enhances the model's ability to identify rare air quality categories correctly. The synthetic samples from SMOTE contribute to better generalization and improved recall for minority classes in subsequent classification tasks.



(a) Class distribution before SMOTE application (b) Class distribution after SMOTE application

Fig. 3. Comparison of class distributions before and after SMOTE application: (a) class imbalance before oversampling; (b) balanced distribution achieved through SMOTE.

Feature engineering plays a critical role in improving model performance. Seasonal features were explicitly incorporated into the LSTM model to enhance its ability to recognize periodic variations in air quality—these engineered attributes aimed to capture cyclical pollutant trends and provide additional context for long-term forecasting. Figures 4 and 5 illustrate that PM2.5 and PM10 concentrations in Hangzhou displayed clear seasonal fluctuations, with higher pollution levels typically occurring in winter.

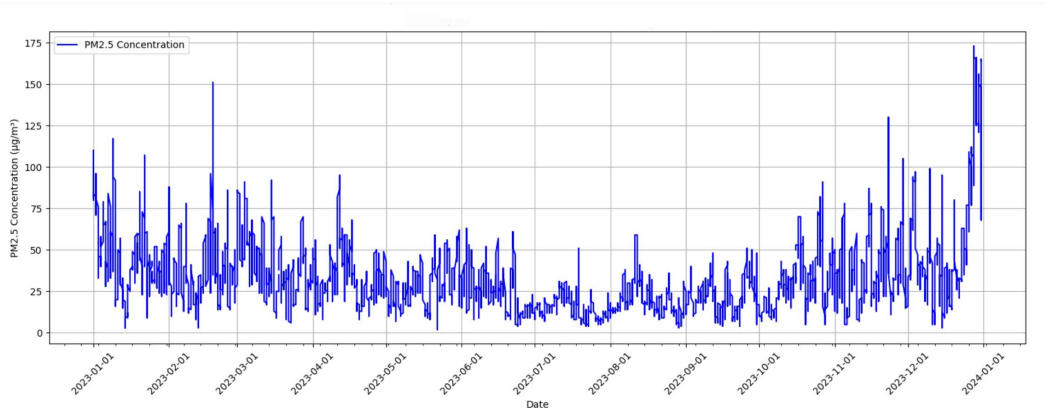


Fig. 4. PM2.5 Concentrations Time Series in 2023 in Hangzhou City

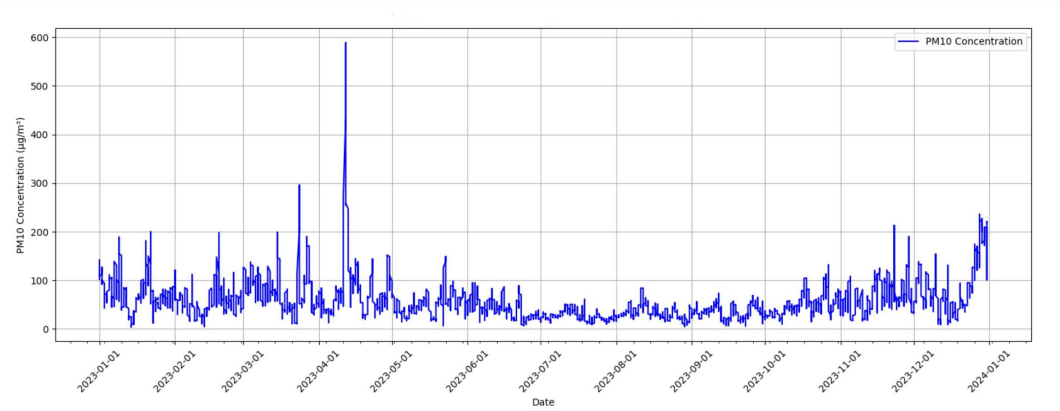


Fig. 5. PM10 concentration time series in 2023 Hangzhou City

The seasonal fluctuations justified the inclusion of seasonal indicators in the LSTM model to help it better model such cyclical patterns. However, experimental results showed that the impact of temporal and seasonal features was not consistently significant across all models and appeared to be model-dependent. For the LSTM model, adding seasonal features and moving averages resulted in modest but consistent improvements [31]. As shown in Table 5, the accuracy increased from 0.9344 to 0.9351, and the macro-PR-AUC improved from 0.8903 to 0.8993, indicating a better precision-recall balance, particularly for minority classes. Additionally, macro and weighted F1-scores experienced slight gains, reinforcing that seasonal attributes contributed to a more balanced classification.

In contrast, ensemble models such as RF and XGBoost exhibited strong predictive performance without notable benefit from explicit seasonal structures. It suggests that engineered temporal features may provide more value in sequential models than tree-based ensembles. Confusion matrices were examined to analyze classification tendencies further to identify misclassification patterns across AQI categories. The results indicate that RF and XGBoost exhibited the lowest misclassification rates, particularly in distinguishing "Moderate" and "Unhealthy" categories, which are often challenging to classify due to overlapping pollutant concentration ranges. In contrast, SVM frequently misclassified "Unhealthy for Sensitive Group" as "Moderate", reflecting its limited capacity to capture subtle variations in pollutant levels. The comparison highlights the importance of

ensemble learning in optimizing classification boundaries in high-dimensional air pollution datasets [9].

The findings of this study are consistent with prior research on air quality classification using machine learning. Previous studies have demonstrated that ensemble methods, particularly Random Forest and XGBoost, outperform traditional classifiers such as SVM in handling complex, nonlinear relationships in air pollution datasets [6]. However, unlike earlier research that heavily relied on meteorological inputs, this study demonstrates that well-engineered pollutant-based features alone can achieve comparable classification accuracy, reinforcing the effectiveness of pollutant-specific modeling approaches. Moreover, the results underscore the role of class imbalance handling techniques such as SMOTE, which has been widely employed in environmental classification tasks to mitigate bias toward majority categories. In summary, Random Forest emerged as the most effective classifier, achieving the highest PR-AUC and F1-score, followed by XGBoost and CatBoost. Feature engineering significantly improved classification accuracy, particularly for models incorporating seasonal and temporal attributes. Class imbalance handling using SMOTE enhanced recall for underrepresented AQI categories but introduced minor trade-offs in precision. The findings suggest that ensemble learning, combined with engineered pollutant-specific features and appropriate imbalance handling, provides a scalable and effective approach for air quality classification. Future work may explore the integration of additional contextual variables, such as meteorological data or traffic emissions, to further enhance predictive accuracy in dynamic urban environments.

4. Conclusions

This study systematically investigated the application of five machine learning models, Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), CatBoost, and Long Short-Term Memory (LSTM) for air quality classification in Hangzhou, utilizing pollutant concentration data. The experimental results demonstrate that ensemble models, particularly RF, perform better in handling high-dimensional environmental data and capturing complex, nonlinear relationships. Random Forest achieved the highest Macro PR AUC (0.9839) and F1-score (0.9727), making it the most effective and robust model in this study. These findings reaffirm the suitability of tree-based ensemble models for structured classification tasks in environmental applications, due to their strong generalization capabilities and resistance to overfitting in imbalanced datasets [6,32]. XGBoost and CatBoost also performed well, demonstrating competitive PR-AUC and F1-scores, benefiting from their gradient boosting frameworks that iteratively refine classification boundaries. Although LSTM was incorporated to leverage temporal dependencies and seasonal attributes, its performance, while improved with feature engineering, remained inferior to ensemble models. The modest improvement in LSTM's accuracy and PR-AUC after introducing seasonal indicators suggests that while temporal features can assist sequential models, they do not substitute for pollutant concentration features, which remain the strongest predictors for urban air quality classification [26]. Data imbalance presented a persistent challenge throughout this study, particularly in predicting rare, extreme pollution events. Implementing SMOTE alleviated some imbalance issues by improving recall for minority classes; however, the trade-off between recall and precision remained evident. This limitation highlights the necessity for future research to explore more advanced imbalance handling techniques, such as cost-sensitive learning or generative data augmentation, to enhance model sensitivity towards underrepresented AQI categories [18]. This study confirms that pollutant-specific feature engineering, combined with ensemble-based models and appropriate class imbalance handling, constitutes a reliable and scalable framework for air quality classification. The Random Forest model (with its tree-based ensemble structure and ability to capture multi-level

feature interactions) emerged as the most robust classifier for the dataset. While deep learning models such as LSTM demonstrated potential for modeling sequential trends, they did not outperform ensemble models in direct classification tasks. Future studies may benefit from incorporating meteorological factors, traffic data, and exploring hybrid architectures such as CNN-LSTM or attention-based models to improve classification performance and generalization across diverse urban environments [33]. Transfer learning techniques and dynamic sampling strategies could also be investigated to enhance predictive capabilities for rare pollution events, making such models more adaptable for real-time air quality monitoring and policy decision-making support.

Acknowledgement
We thank the Center for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, for academic guidance and support throughout this study.

References

[1] Dominski, F. H.; Lorenzetti Branco, J. H.; Buonanno, G.; Stabile, L.; Gameiro da Silva, M.; Andrade, A. *Effects of Air Pollution on Health: A Mapping Review of Systematic Reviews and Meta-Analyses*. Environmental Research 2021, 201, 111487. DOI:10.1016/j.envres.2021.111487.

[2] Guo, Z.; Wang, X.; Ge, L. *Classification Prediction Model of Indoor PM2.5 Concentration Using CatBoost Algorithm*. Frontiers in Built Environment 2023, 9. DOI: 10.3389/fbuil.2023.1207193.4

[3] Zhao, Y.; Zhang, X.; Chen, M.; Gao, S.; Li, R. *Regional Variation of Urban Air Quality in China and Its Dominant Factors.* Journal of Geographical Sciences 2022, 32, 853–872. DOI: 10.1007/s11442-022-1975-8.

[4] Feng, R.; Wang, Q.; Huang, C.; Liang, J.; Luo, K.; Fan, J.; Cen, K. *Investigation on Air Pollution Control Strategy in Hangzhou for Post-G20/Pre-Asian Games Period (2018–2020).* Atmospheric Pollution Research 2019, 10, 197–208. DOI: 10.1016/j.apr.2018.07.006.

[5] Chadalavada, S.; Faust, O.; Salvi, M.; Seoni, S.; Raj, N.; Raghavendra, U.; Gudigar, A.; Barua, P. D.; Molinari, F.; Acharya, R. *Application of Artificial Intelligence in Air Pollution Monitoring and Forecasting: A Systematic Review*. Environmental Modelling & Software 2025, 185, 106312. DOI: 10.1016/j.envsoft.2024.106312.

[6] Kumar, K.; Pande, B. P. *Air Pollution Prediction with Machine Learning: A Case Study of Indian Cities.* International Journal of Environmental Science and Technology 2022, 20, 5333–5348. DOI: 10.1007/s13762022-04241-5.

[7] Guo, Y.; Li, K.; Zhao, B.; Shen, J.; Bloss, W. J.; Azzi, M.; Zhang, Y. *Evaluating the Real Changes of Air Quality Due to Clean Air Actions Using a Machine Learning Technique: Results from 12 Chinese Mega-Cities During 2013–2020.* Chemosphere 2022, 300, 134608. DOI: 10.1016/j.chemosphere.2022.134608.

[8] Lakshmi, V. S.; Vijaya, M. S. *Feature Selection Techniques for Building Robust Air Quality Prediction Model*. Algorithms for Intelligent Systems 2024, 185–198. DOI: 10.1007/978-981-99-9436-613.

[9] Lei, T. M. T.; Ng, S. C. W.; Siu, S. W. I. *Application of ANN, XGBoost, and Other ML Methods to Forecast Air Quality in Macau*. Sustainability 2023, 15, 5341. DOI: 10.3390/su15065341

[10] Liao, Q.; Zhu, M.; Wu, L.; Pan, X.; Tang, X.; Wang, Z. *Deep Learning for Air Quality Forecasts: A Review*. Current Pollution Reports 2020, 6, 399–409. DOI: 10.1007/s40726-020-00159-z.

[11] Zhu, D.; Cai, C.; Yang, T.; Zhou, X. (2018). *A machine learning approach for air quality prediction: Model regularization and optimization*. Big Data and Cognitive Computing 2(1), 5. DOI: 10.3390/bdcc2010005.

[12] Sabir, M. F.; Ragab, M.; Khadidos, A. O.; Alyoubi, K. H.; Khadidos, A. O. (2024). *Ensemble deep learning-based air pollution prediction for sustainable smart cities*. Computer Systems Science and Engineering 48, 627–643. DOI: 10.32604/csse.2023.041551.

[13] Liu, W.; Wu, J.; Liu, X.; et al. (2022). *Time-sensitive prediction of NO2 concentration in China using an ensemble machine learning approach*. Ecological Indicators 141, 109121. DOI: 10.1016/j.ecolind.2022.109121.

[14] Liang, Y.-C.; Maimury, Y.; Chen, A. H.-L.; Juarez, J. R. C. *Machine Learning-Based Prediction of Air Quality*. Applied Sciences 2020, 10, 9151. DOI: 10.3390/app10249151.

[15] Wang, J.; Xu, Z.; Liu, L.; et al. *Time-Sensitive Prediction of NO2 Concentration in China Using an Ensemble Machine Learning Model*. Ecological Indicators 2022, 141, 109121. DOI: 10.1016/j.ecolind.2022.109121.

[16] Chandra, W.; Suprihatin, B.; Resti, Y. *Median-KNN Regressor-SMOTE-Tomek Links for Handling Missing and Imbalanced Data in Air Quality Prediction.* Symmetry 2023, 15, 887. DOI: 10.3390/sym15040887.

[17] Zaini, N.; Ean, L. W.; Ahmed, A. N.; Abdul Malek, M.; Chow, M. F. *PM2.5 Forecasting for an Urban Area Based on Deep Learning and Decomposition Method.* Scientific Reports 2022, 12, Article 21769. DOI: 10.1038/s41598-022-21769-1.

[18] Jayadi, B. V.; Lauro, M. D.; Rusdi, Z.; Handhayani, T. *Air Quality Index Classification for Imbalanced Data Using Machine Learning Approach*. Sistemasi 2024, 13, 951–951. DOI: 10.32520/stmsi.v13i3.3503.

[19] Zhang, Z.; Zhang, S.; Chen, C.; Yuan, J. *A Systematic Survey of Air Quality Prediction Based on Deep Learning*. Alexandria Engineering Journal 2024, 93, 128–141. DOI: 10.1016/j.aej.2024.03.031

[20] Probst, P.; Wright, M. N.; Boulesteix, A.-L. *Hyperparameters and Tuning Strategies for Random Forest*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2019, 9. DOI: 10.1002/widm.1301. Available online: https://arxiv.org/abs/1804.03515.

[21] Chen, T.; Guestrin, C. *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining' DD' 16) 2016, 785–794. DOI: 10.1145/2939672.2939785

[21] Hancock, J. T.; Khoshgoftaar, T. M. *CatBoost for Big Data: An Interdisciplinary Review*. Journal of Big Data 2020, 7. DOI:10.1186/s40537-020-00369-8.

[22] Cortes, C.; Vapnik, V. *Support-Vector Networks*. Machine Learning 1995, 20, 273–297. DOI:10.1007/BF00994018.

[23] Liu, B.; Jin, Y.; Li, C. *Analysis and Prediction of Air Quality in Nanjing from Autumn 2018 to Summer 2019 Using PCR–SVR–ARMA Combined Model.* Scientific Reports 2021, 11, Article 101. DOI: 10.1038/s41598-020-79462-0

[24] Hossin, M.; Sulaiman, M. N. *A Review on Evaluation Metrics for Data Classification Evaluations*. International Journal of Data Mining & Knowledge Management Process (IJDKP) 2015, 5(2), 1–11. DOI: 10.5281/zenodo.3557376.