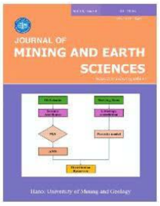
Journal of Mining and Earth Sciences

Website: https://jmes.humg.edu.vn



RF-PM$_{10}$Hybrid model for real-time PM$_{10}$ forecasting in open-pit copper mine: A case study at the Sin Quyen Copper Mine

Ngoc Tuan Le[1], Hoang Nguyen[2,3]*, Xuan-Nam Bui[3,4], Hoa Thu Thi Le[2,3]

[1] Vinacomin – Minerals Holding Corporation, Hanoi, Vietnam
[2] Department of Surface Mining, Hanoi University of Mining and Geology, Hanoi, Vietnam
[3] Innovations for Sustainable and Responsible Mining (ISRM) Research Group, Hanoi University of Mining and Geology, Hanoi, Vietnam
[4] Vietnam Mining Science and Technology Association, Hanoi, Vietnam

ARTICLE INFO

Article history:
Received 25th Mar. 2025
Revised 10th May 2025
Accepted 27th May. 2025

Keywords:
Air Quality Prediction,
Environmental Monitoring,
Feature Engineering,
Machine Learning,
Open-Pit Mining,
PM$_{10}$ Forecasting,
Principal Component Analysis (PCA),
RF-PM10Hybrid,
SMOGN.

ABSTRACT

Air pollution in open-pit mining areas poses significant environmental and health risks, with particulate matter (PM$_{10}$) being one of the most critical pollutants. Accurate forecasting of PM$_{10}$ concentrations is essential for real-time air quality management and dust mitigation strategies. This study develops a machine learning-based framework for PM$_{10}$ prediction at the Sin Quyen open-pit copper mine, leveraging advanced feature engineering, Principal Component Analysis (PCA), and Synthetic Minority Over-sampling technique for Regression (SMOGN) to enhance model accuracy. Six forecasting models were evaluated, including Random Forest (RF-PM$_{10}$Hybrid), XGBoost, LightGBM, ARIMA, SARIMA, and Holt-Winters exponential smoothing. The results indicate that machine learning models significantly outperform traditional time-series models with RMSE of 5.791, 8.293, 6.172, 4.233, 11.070, 13.108; MAE of 3.518, 3.953, 3.770, 4.208, 8.800, 10.224; MAPE of 11.70%, 13.18%, 12.57%, 14.03%, 29.32%, 34.07% for the RF-PM10Hybrid, XGBoost, LightGBM, ARIMA, SARIMA, Holt-Winters, respectively. RF-PM$_{10}$Hybrid achieved the best forecasting performance, with the lowest RMSE (5.791) and MAE (3.518) on the testing dataset, followed by LightGBM and XGBoost. Conversely, statistical models (ARIMA, SARIMA, and Holt-Winters) exhibited higher forecasting errors, making them less suitable for predicting PM$_{10}$ variations in open-pit mining environments. Key methodological advancements include the integration of lag features, rolling statistics, and interaction terms, which improved the ability of ML models to capture PM$_{10}$ dynamics. SMOGN was applied to balance the dataset, ensuring better representation of high- PM$_{10}$ events. The findings demonstrated that machine learning-based approaches, particularly RF-PM10Hybrid, provide a reliable tool for real-time PM$_{10}$ forecasting, supporting proactive dust control, regulatory compliance, and sustainable mining operations.

**Corresponding author*
E - mail: nguyenhoang@humg.edu.vn

DOI: 10.46326/JMES.2025.66(4).04

1. Introduction

Open-pit mining operations generate significant amounts of airborne pollutants, particularly particulate matter (PM), due to activities such as drilling, blasting, material handling, transportation, and excavation. Among these pollutants, PM_{10} (particulate matter ≤10 μm) is of particular concern due to its ability to remain suspended in the atmosphere for extended periods and travel over long distances, which can penetrate the respiratory system and cause significant health and environmental impacts. Unlike larger particles, PM_{10} can penetrate deep into the human respiratory system, leading to severe respiratory and cardiovascular diseases. According to the national air quality standard QCVN 05:2023/BTNMT (Vietnam), the permissible annual average concentration of PM_{10} is 50 $\mu g/m^3$, and the 24-hour average permissible limit is 150 $\mu g/m^3$.

From an environmental perspective, PM_{10} pollution contributes to atmospheric haze, soil contamination, and water pollution through particle deposition. Fine dust particles can accumulate in surrounding agricultural lands, reducing crop productivity and altering soil composition. Additionally, PM_{10} emissions affect local ecosystems by degrading air quality and disrupting wildlife in mining regions.

The primary sources of PM_{10} in open-pit mining operations, including blasting activities, material handling, haul truck movements, wind erosion from exposed surfaces, and crushing processes. For mine operations, excessive PM_{10} levels can lead to regulatory fines, reduced worker productivity, increased equipment maintenance costs, and reputational damage. Many countries have imposed strict air quality regulations requiring mines to monitor and control dust emissions effectively. Failure to comply with these standards can result in mine shutdowns or production halts, significantly impacting operational efficiency and profitability.

Given the severe consequences of PM_{10} pollution, real-time air quality forecasting has become a critical tool for mine operators to mitigate health risks, enhance environmental compliance, and improve operational efficiency.

Traditional dust control strategies react to high PM_{10} levels only after they exceed regulatory limits. However, real-time forecasting enables a proactive approach, allowing mine operators to anticipate dust concentration trends and take preventive actions before levels become hazardous. The importance of PM_{10} forecasting is timely prediction of PM_{10} concentrations allows mine operators to implement dust suppression measures proactively (e.g., water spraying, operational scheduling), contributing to environmental compliance, worker health protection, and sustainable mining operations.

In this study, we develop and evaluate a hybrid machine learning model (RF-PM_{10}Hybrid) to improve PM_{10} forecasting in open-pit mining. By integrating feature engineering, dimensionality reduction (PCA), and imbalanced data handling (SMOGN) into Random Forest, this model enhances forecasting accuracy and robustness. A real-world case study is conducted at Sin Quyen Copper Mine, Vietnam, to validate the model's effectiveness in improving real-time PM_{10} predictions and supporting data-driven mine management strategies.

Traditional time-series models such as AutoRegressive Integrated Moving Average (ARIMA) and Seasonal AutoRegressive Integrated Moving Average (SARIMA) have been widely used for air pollution forecasting due to their ability to capture temporal dependencies and seasonal patterns. However, when applied to complex, dynamic environments like open-pit mines, these models exhibit several critical limitations:

a) Inability to capture non-linear relationships

- PM_{10} concentration is influenced by multiple meteorological and operational factors, including wind speed, temperature, humidity, atmospheric pressure, and mining activities.

- ARIMA and SARIMA assume linear relationships in the data, making them ineffective for capturing non-linear interactions between these factors.

b) Poor handling of sudden variations in PM_{10} levels

- Blasting, excavation, and haul truck movements can cause sharp spikes in PM_{10} levels, which traditional models fail to predict accurately.

- SARIMA can model seasonal variations, but it struggles to adapt to sudden short-term fluctuations, leading to high forecasting errors during critical dust events.

c) Limited performance with small, noisy datasets

- Time-series models require stationary data (constant mean and variance), but PM_{10} levels fluctuate significantly due to changing weather conditions and mining schedules.

- Data pre-processing techniques such as differencing can help stabilize variance, but they also remove valuable trend information, affecting prediction accuracy.

d) Lack of multi-feature learning capability

- ARIMA and SARIMA rely only on past values of PM_{10} (univariate approach), ignoring external meteorological and operational data.

- A multi-feature approach (incorporating meteorological factors) is necessary to improve forecasting accuracy, but traditional models lack this capability.

To overcome these limitations, hybrid machine learning (ML) models have emerged as a powerful alternative for PM_{10} forecasting in open-pit mining. The combination of statistical, feature engineering, and ML-based approaches provides significant advantages over traditional models, specifically:

a) Machine learning models can capture complex relationships

- ML models such as Random Forest (RF), XGBoost, and LightGBM can learn from multiple meteorological and operational factors, capturing both linear and non-linear dependencies.

- Feature engineering techniques (e.g., lag features, rolling statistics, interaction terms) further enhance ML model interpretability and accuracy.

b) Hybrid models leverage the strengths of both ML and time-series approaches

- Traditional models (e.g., SARIMA) handle seasonal trends well but fail with complex feature interactions.

- ML models learn multi-feature dependencies but struggle with sequential data patterns.

- A hybrid approach combines the best of both worlds, ensuring accurate short-term and long-term PM_{10} forecasts.

c) Imbalanced data handling with SMOGN improves prediction of extreme PM_{10} events

- Extreme PM_{10} events (high pollution peaks) are rare but critical for mine safety and regulation compliance.

- Traditional models often underpredict extreme values due to data imbalance.

- SMOGN (Synthetic Minority Over-sampling for Regression) generates synthetic samples to improve model learning on rare events, ensuring better performance in forecasting high PM_{10} levels.

d) Dimensionality reduction (PCA) enhances model efficiency

- PM_{10} forecasting models require multiple input variables, but too many features can lead to overfitting and increased computational cost.

- Principal Component Analysis (PCA) reduces feature dimensionality while retaining important variance, optimizing model performance.

Thus, this study introduces RF-PM_{10}Hybrid, a hybrid Random Forest-based model that integrates advanced feature engineering, PCA, and SMOGN to overcome the limitations of traditional forecasting methods. The proposed approach is evaluated through a case study at Sin Quyen Copper Mine, Vietnam, demonstrating its effectiveness in real-time air quality forecasting. The key contributions of our study include:

- RF-PM_{10}Hybrid model: Integrates RF with feature engineering, PCA, and SMOGN for PM_{10} forecasting.

- Advanced feature engineering: Includes lag features, rolling statistics, and interaction terms.

- Dimensionality reduction (PCA): Reduces computational cost while maintaining 90% variance.

- Handling data imbalance (SMOGN): Improves prediction of extreme PM_{10} levels.

- Comparative model evaluation: Benchmarks RF-PM_{10}Hybrid against XGBoost, LightGBM, ARIMA, SARIMA, Holt-Winters.

- Real-world application: Validates model performance at Sin Quyen Copper Mine, Vietnam.

2. Literature review and principle of machine learning models

2.1. Literature review

Air pollution forecasting, particularly PM_{10} prediction, has been extensively studied using both traditional time-series models and machine learning approaches. Each method has its strengths and limitations, leading to the need for hybrid models that leverage the advantages of both.

For forecasting PM10, Sumanth et al. (2020) modelled PM_{10} dispersion in coal mines using the AERMOD model and evaluates the effects of different digital elevation models (DEMs) on dispersion predictions. The authors compared multiple DEMs (SRTM, ASTER, CartoDEM) for terrain representation and assessed their effects on PM_{10} dispersion predictions. They evaluated the performance of AERMOD using Willmott's Index of Agreement and other performance metrics. The study found that overburden dumps, haulage routes, and railway sidings contributed the most to PM_{10} emissions. Using more recent DEMs improved model performance.

In another study, Sánchez Lasheras et al. (2020) compared different machine learning models for forecasting PM_{10} concentrations in a port area using historical air quality data. The authors used various models, including ARIMA, Vector Autoregressive Moving Average (VARMA), Multilayer Perceptron (MLP), Support Vector Machines for Regression (SVMR), and Multivariate Adaptive Regression Splines (MARS). The best short-term forecasts (1 month ahead) were achieved using SVMR, while MLP performed best for longer-term forecasts (6 months ahead). ARIMA was found to be less effective than ML models for long-term forecasting. This paper provides strong evidence that ML models outperform traditional statistical methods for air quality forecasting.

Török et al. (2023) also applied various machine learning models, including Random Forest, Gradient Boosting, and Neural Networks, to predict PM_{10} concentrations in an industrial region. They found that Gradient Boosting models performed best, highlighting the importance of ensemble learning. Feature importance analysis showed that meteorological factors significantly impact PM_{10} levels.

Time-series forecasting models such as AutoRegressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and Holt-Winters Exponential Smoothing have been widely used for predicting $PM_{2.5}$ and PM10 concentrations (Pozza et al., 2010; Bhatti et al., 2021; da Silva et al., 2023). These methods rely solely on historical PM_{10} values to identify temporal patterns and project future trends.

In Vietnam, the importance of air quality monitoring and dust control in open-pit mines has also been recognized. In 2018, a research team from Hanoi University of Mining and Geology (HUMG) led by Prof. Dr. Bui Xuan Nam collaborated with Dong-A University (Korea) to develop an air quality control system for deep open-pit coal mines in Quang Ninh Province, under the sponsorship of a bilateral international cooperation research project funded by the Ministry of Education and Training (Bui, 2021). The study selected three representative mines-Deo Nai, Cao Son, and Coc Sau-for assessing air quality and proposing effective dust control measures. This work laid the foundation for air quality monitoring and management in deep open-pit mines in Vietnam, yielding promising and practical results. These early initiatives have highlighted the necessity for developing predictive models to support real-time dust control strategies in Vietnamese mining operations. However, this study has not applied AI-based models for forecasting air quality in these mines. Furthermore, the air quality in the open-pit coal mines are different from the open-pit copper mine.

To address the limitations of traditional models, machine learning (ML) methods such as Random Forest (RF), XGBoost, and LightGBM have been widely adopted for PM_{10} forecasting. These models offer higher accuracy and adaptability by learning complex relationships between multiple meteorological and environmental factors.

2.2. Random Forest (RF)

Random Forest is an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting (Fratello and Tagliaferri, 2018; Halabaku and Bytyçi, 2024). It works by bootstrapping with randomly selecting subsets of data for training multiple trees. Then, it applies feature splitting and finding optimal split points for decision-making. Finally, it conducts aggregation (voting for classification/averaging for regression) by combining predictions from multiple trees to improve generalization.

The main advantages of RF include handles non-linear dependencies between PM_{10} and meteorological factors (temperature, humidity, wind speed, etc.), works well with high-dimensional data and selects important features automatically, and robust against overfitting due to averaging across multiple trees.

However, the limitations of RF including slower in training compared to single-tree models and feature importance may be biased toward variables with more levels (e.g., categorical variables) (Hegelich, 2016).

Why RF is chosen as the core model in RF-PM_{10}Hybrid? Because RF is highly interpretable, making it suitable for environmental applications (Wang et al., 2021; Simon et al., 2023). Moreover, it can be easily combined with PCA and SMOGN to improve forecasting accuracy. It also performs well with multi-variable inputs (meteorological and operational features).

2.3. XGBoost (Extreme Gradient Boosting)

XGBoost is a gradient boosting algorithm that iteratively improves prediction accuracy by training decision trees sequentially, minimizing loss using gradient descent optimization, and applying regularization (L1 & L2) to prevent overfitting (Asselman et al., 2023; Sibindi et al., 2023; Kavitha and Priyadharshini, 2024).

The main advantages of XGBoost including fast and optimized for large datasets, effective in handling missing values and imbalanced data, and better generalization compared to individual decision trees.

Nevertheless, the limitations of XGBoost including high computational cost compared to simpler models and sensitive to hyperparameter tuning, requiring optimization for best results.

2.4. LightGBM (Light Gradient Boosting Machine)

LightGBM is an optimized version of XGBoost, designed for faster training and lower memory usage. Instead of growing trees depth-first, LightGBM splits leaf nodes first, allowing faster training speeds, making it ideal for real-time applications, handling large datasets with minimal memory consumption, and efficient feature selection, focusing on the most important variables (Zhang and Gong, 2020, Wang et al., 2025).

Main advantages of LightGBM include up to 10x faster than XGBoost while maintaining similar accuracy and performs well on imbalanced datasets with appropriate tuning.

However, the limitations of LightGBM include more prone to overfitting compared to Random Forest and less interpretable than RF, making it harder to explain model decisions.

2.5. Limitations and solutions in this study

Despite significant advancements in PM_{10} forecasting using both traditional time-series models and machine learning methods, several challenges remain when applying these approaches in open-pit mining environments. These challenges include ineffective handling of seasonal variations, difficulty in capturing temporal dependencies, and bias due to class imbalance in PM_{10} data.

In open-pit mining, PM_{10} concentrations are influenced by cyclic operational and meteorological patterns, including daily variations in mining activity levels (e.g., shifts, blasting schedules), seasonal weather effects (e.g., winter humidity reducing dust, dry summers increasing airborne particles), and wind direction changes across different seasons affecting dust dispersion.

In addition, time-series forecasting requires understanding the relationships between past PM_{10} values and future trends. However, standard machine learning models like RF, XGBoost, and LightGBM are not inherently designed for sequential data.

Furthermore, PM_{10} data in mining environments is often highly imbalanced, with long periods of low pollution levels and infrequent but critical high-concentration events (e.g., post-blasting or dry season spikes). The limitations and solutions in this study are summarized in Table 1.

3. Methodology

3.1. Study area and dataset

The Sin Quyen Copper Mine, located in Lao Cai Province, Vietnam, is one of the largest open-pit copper mines in the region. Operated by Vietnam National Coal and Mineral Industries Group (VINACOMIN), the mine produces significant amounts of dust due to blasting, excavation, material transportation, and ore processing activities.

To monitor air quality and environmental conditions, an air quality monitoring system was deployed and strategically placed at +102 m to capture variations in PM_{10} levels influenced by mining operations and meteorological conditions. The system continuously recorded particulate matter (i.e., PM_{10}) concentrations along with key meteorological variables. It should be noted that due to the impacts of all operations in the mine, the surrounding air quality may be affected, and therefore, we selected this place (i.e., +102 m) on the surface of the mine to evaluate the pollution of PM_{10} on the surrounding environment.

The dataset used in this study was collected from April 10, 2024, to September 30, 2024, covering a period of nearly six months. During this time, measurements were taken at different timestamps (1 min, 2 min, 3 min, and 4 min intervals), resulting in a high-resolution dataset suitable for short-term PM_{10} forecasting. The raw data underwent preprocessing and resampling to ensure uniform hourly intervals, facilitating effective model training and evaluation.

In this study, the PM_{10} forecasting models will be developed based on data from a fixed air quality monitoring station installed at a representative location within the Sin Quyen open-pit copper mine. Although only one station was used, it captures the major dust emissions and air quality variations related to mining operations across the site.

The dataset consists of six key variables, including meteorological factors and PM_{10} concentrations, which are critical for accurate air quality forecasting. These variables are summarized in Table 2.

Table 1. Summary of limitations & solutions in this study.

Limitation	Issues in existing models	Proposed solutions
Seasonal trends poorly captured	ARIMA, SARIMA struggle with irregular mining activity	Time-based features + rolling statistics in RF-PM_{10}Hybrid
Static ML models fail to learn time dependencies	RF, XGBoost, LightGBM treat data as independent observations	Adding lag features + rolling statistics to ML models
Imbalanced PM_{10} distribution leads to bias	ML models underpredict extreme PM_{10} levels	Applying SMOGN for synthetic minority oversampling

Table 2. Description of variables used in forecasting.

Variable	Unit	Description
Humidity	%	Relative humidity of the air, affecting dust suspension and dispersion.
Temperature	°C	Ambient air temperature, influencing dust settling rates.
Pressure	Pa	Atmospheric pressure, affecting air density and dust transport.
Wind Direction	Degrees (°)	Direction from which the wind is blowing, influencing dust movement.
Wind Speed	m/s	Speed of the wind, impacting the rate of PM_{10} dispersion.
PM_{10} (dust10)	$\mu g/m^3$	Concentration of particulate matter ≤10 μm in diameter, representing air quality.

Table 3. Summary of statistical indexes of the dataset.

	Humidity	Temperature	Pressure	Wind direction	Wind speed	PM_{10}
count	137761	137761	137761	137761	137761	137761
mean	72.110	29.334	99836.455	59.774	2.524	37.541
std	16.043	9.161	501.567	52.687	2.938	547.185
min	21	17.5	98373	0	0	-32768
25%	60	18.204	99421	11.42	0.71	9
50%	74	32	99859	47.446	1.29	30
75%	88.798	35.8	100329.6	102	3.421	41
max	99	50.1	100741	255	37.23	32000

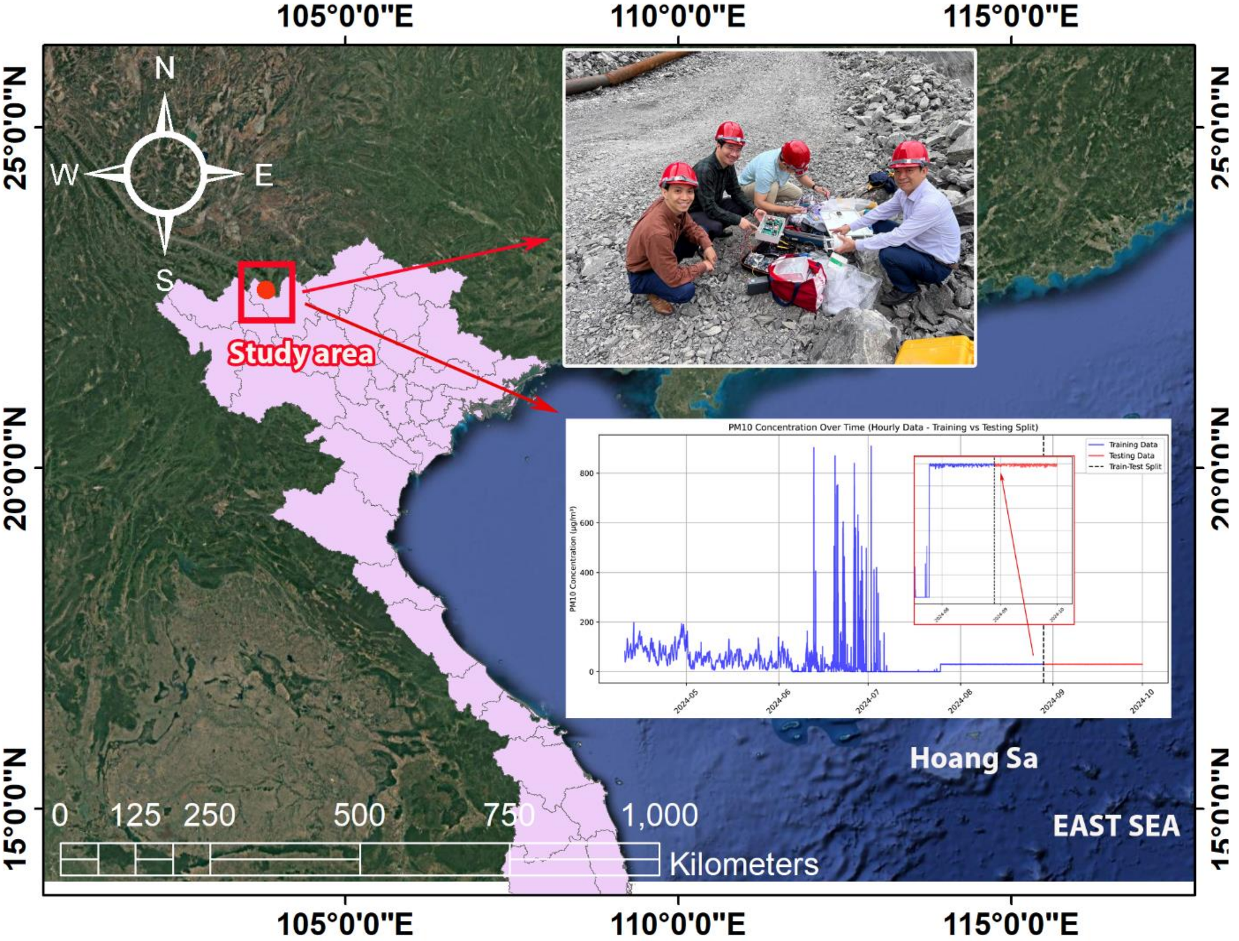


Figure 1. Location of study area and hourly intervals of PM_{10} collected in this study.

These variables play a crucial role in PM_{10} concentration forecasting, as dust dispersion is highly dependent on meteorological conditions. Wind speed and direction significantly impact the spread and accumulation of dust particles, while humidity and temperature influence the rate of dust suspension and deposition. Table 3 presents the statistical indexes of the dataset collected in this study.

Given the variability in time intervals in the raw dataset, all data was resampled to hourly intervals using linear interpolation, as shown in Figure 1. This step ensures a consistent temporal resolution, making it suitable for training the hybrid machine learning models.

3.2. Data preprocessing and feature engineering

Effective data preprocessing and feature engineering are crucial for improving the accuracy and robustness of PM_{10} forecasting models. This study employs multiple preprocessing techniques to handle missing values, create lag features, capture short-term trends, and generate interaction terms that enhance the predictive power of the RF-PM_{10}Hybrid model.

3.2.1. Handling negative values

Based on the summary statistics of the dataset in Table 3, we can see that the average humidity is 72.1%, ranging from 21% to 100%; The mean temperature is 29.3^{0}C, with a minimum of 17.5^{0}C; The average atmospheric pressure is 99836 Pa, with a minimum value of 98373 Pa; The mean wind direction is 59.77°, spanning from 0^0 to 360^0, and the average wind speed is 2.52 m/s, with some instances showing 0 m/s. Critical observations for PM_{10} concentration were also observed with nean PM_{10} concentration of 37.54 μg/m^3, which appears reasonable. The standard deviation of 547.18 μg/m^3, indicating a very high level of dispersion in the data. Minimum value of -32,768 μg/m^3, which is clearly an erroneous value that needs to be corrected. The 25th percentile (Q1) of 9 μg/m^3, meaning 25% of the data has PM_{10} concentrations lower than this threshold, and finally, the maximum value is extremely high, suggesting possible outliers due to mining explosions.

Remarkably, the massive standard deviation and the presence of negative values (-32,768) suggest sensor errors, data corruption, or incorrect data entries. Therefore, the handling negative values should be handled before developing AI-based models for forecasting PM_{10} in this study.

After checking the dataset, the number of negative PM_{10} values is 38 in the whole dataset, as shown in Figure 2. These values may be due to sensor errors, thus these negative values were replaced to zero, and the dataset after handling negative values is shown in Figure 3.

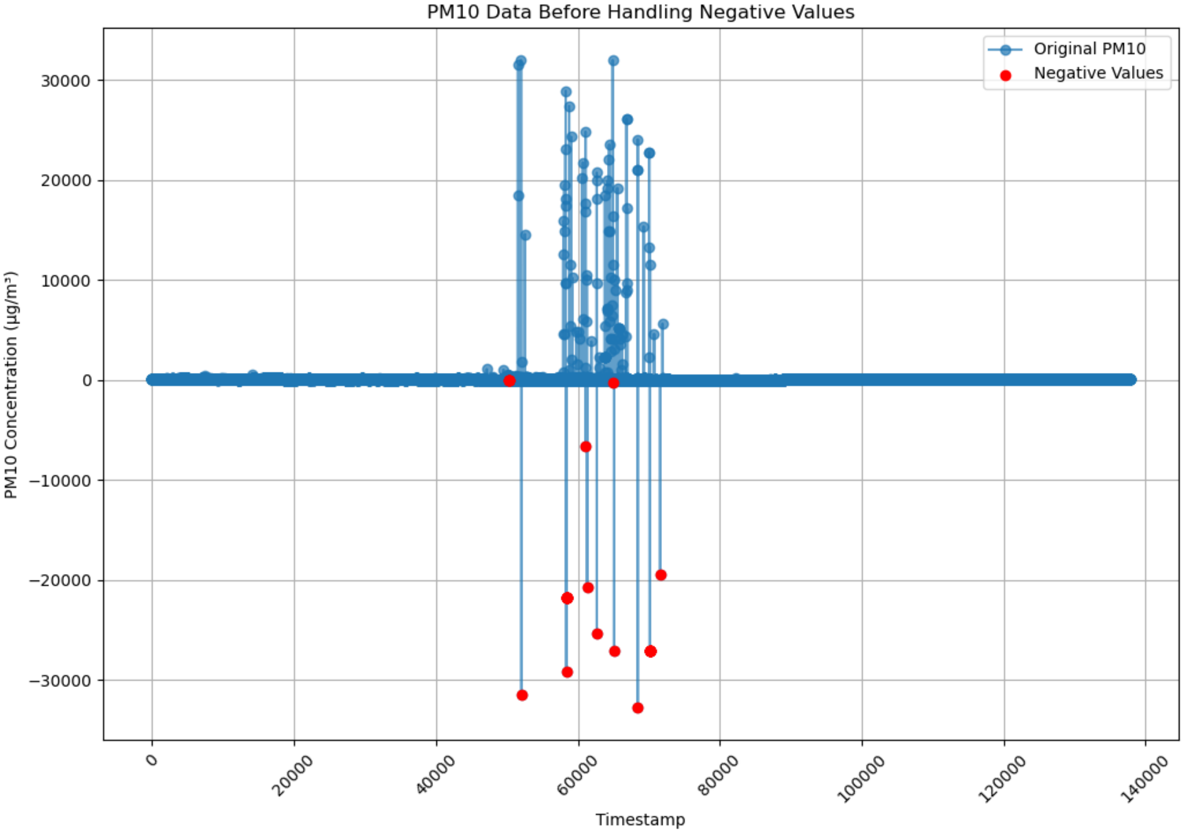


Figure 2. Negative PM_{10} values before handling.

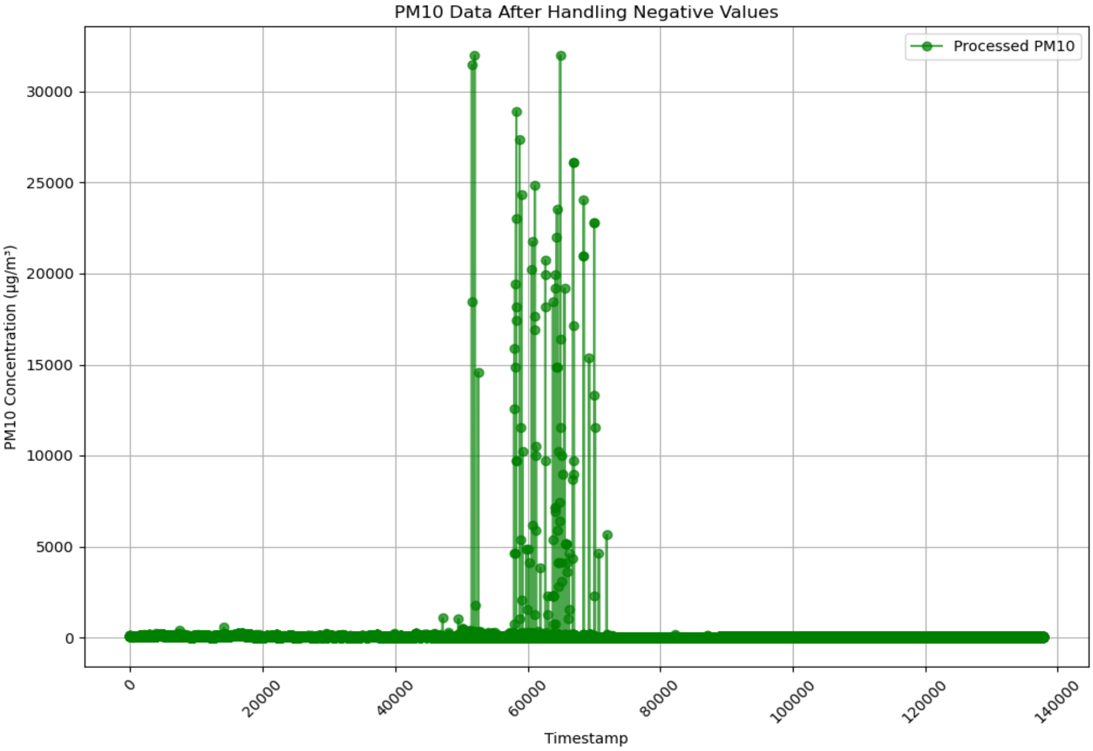


Figure 3. Negative PM_{10} values after handling.

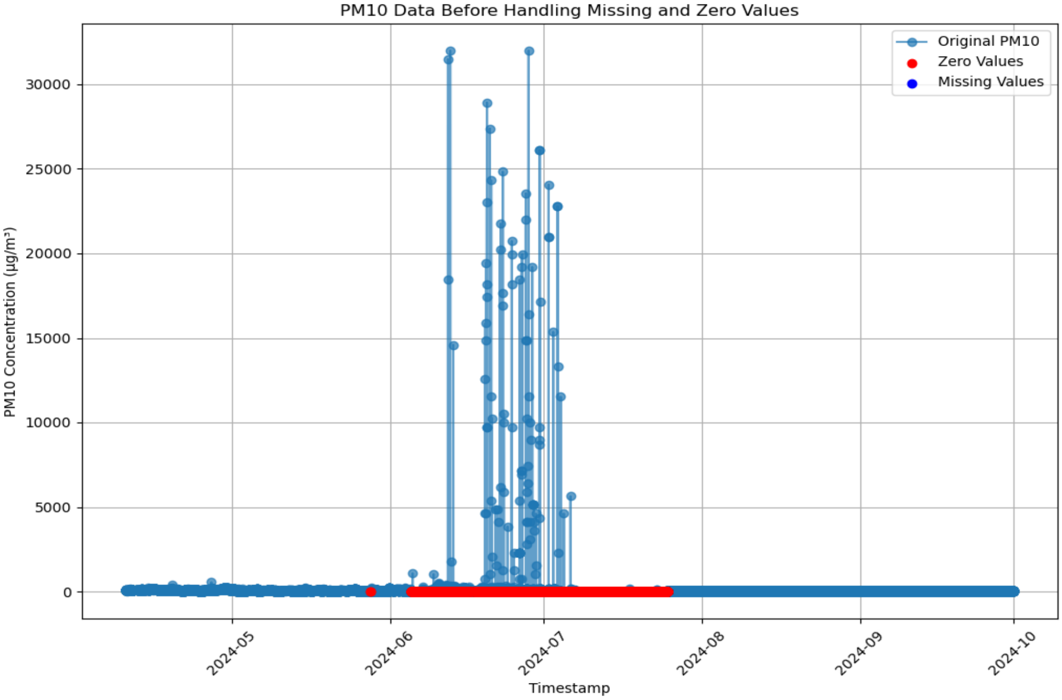


Figure 4. PM_{10} data before handling missing and zero values.

3.2.2. Handling missing values and replacing zero values with the median

The raw dataset contains PM_{10} (dust10) and meteorological variables collected at varying time intervals (1 min, 2 min, 3 min, and 4 min). Before model training, the dataset was resampled to hourly intervals using linear interpolation to ensure uniformity, as shown in Figure 4.

To resampling to hourly intervals, linear interpolation was used to fill missing timestamps.

For missing meteorological data, missing values in humidity, temperature, pressure, wind direction, and wind speed were filled using nearest neighbor interpolation.

For missing PM_{10} values, missing dust concentrations were filled using the median value of the surrounding time steps to avoid extreme fluctuations.

In mining environments, zero PM_{10} values are unrealistic and typically result from sensor errors or faulty data transmission. Zero PM_{10} values were replaced with the median of the dataset to maintain data integrity without introducing bias. Finally, zero values in dust10 replaced with the median of the entire dataset, and non-zero missing values were filled using linear and nearest-neighbor interpolation.

By addressing missing and zero values, the dataset was cleaned and prepared for feature extraction, ensuring that the model learned from accurate and representative data, as shown in Figure 5.

3.2.3. Generating lag features (PM_{10} values from previous hours)

PM_{10} levels are strongly influenced by historical concentrations, as pollution trends follow temporal patterns. To help the model learn these dependencies, lag features were introduced.

In this study, lag features are important, as PM_{10} levels at a given hour depend on previous concentrations, and including past PM10 values as input features allows the model to recognize short-term trends.

For each time step (t), the following lag features were generated: lag_1: PM_{10} concentration at t-1 hour. lag_2: PM_{10} concentration at t-2 hours. lag_3: PM_{10} concentration at t-3 hours. lag_4: PM_{10} concentration at t-4 hours. lag_5: PM_{10} concentration at t-5 hours. lag_6: PM_{10} concentration at t-6 hours. These lagged features enable the model to identify temporal dependencies, making it more effective in capturing trends in PM_{10} levels.

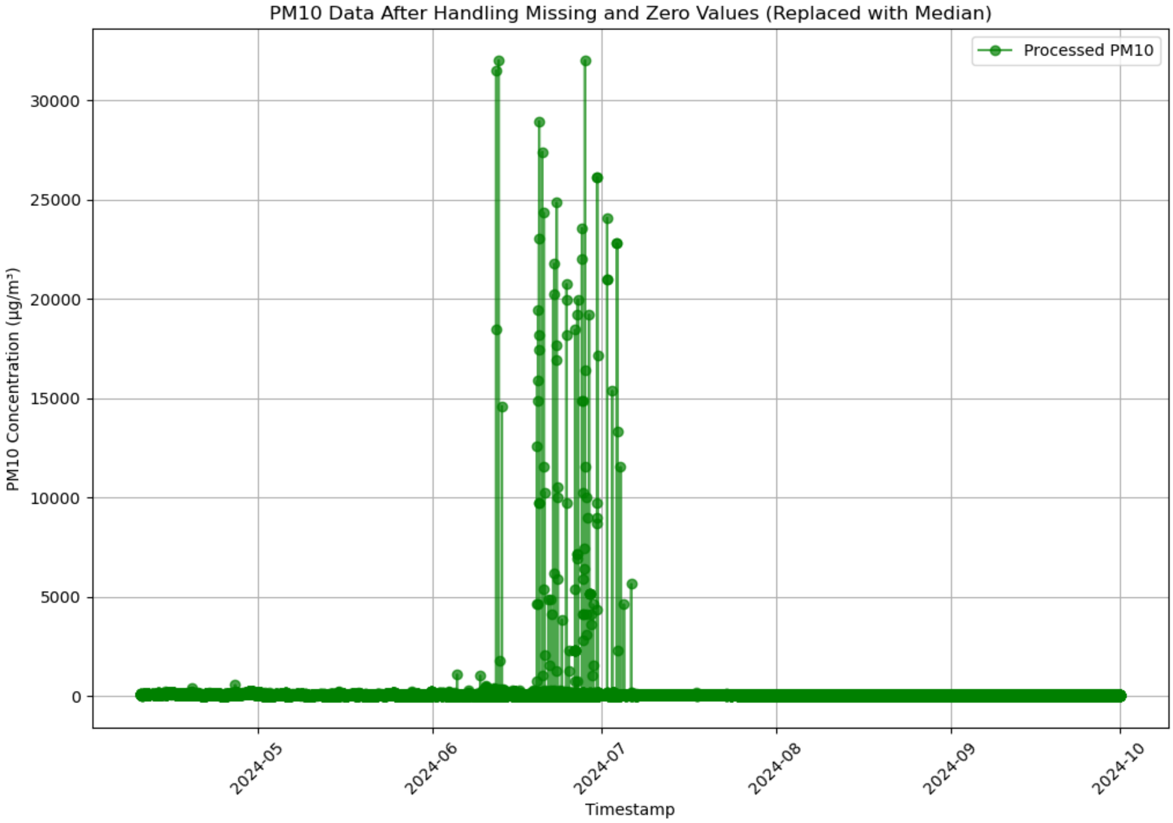


Figure 5. PM_{10} data after handling missing and zero values.

3.2.4. Rolling statistics (mean, standard deviation) to capture short-term trends

In addition to lag features, rolling window statistics were used to model short-term variations in PM_{10} levels, because PM_{10} fluctuations occur due to changes in wind speed, humidity, and mining activities, averaging over short time periods reduces noise and improves trend identification, and standard deviation measures volatility, helping the model detect periods of sudden PM_{10} increases.

For each time step (t), the following rolling statistics were computed: Rolling mean (5-hour window): Average PM_{10} levels over the past 5 hours. Rolling standard deviation (5-hour window): Variability of PM_{10} over the past 5 hours. These strategies are calculated using Eqs. (1) and (2) as follow.

$$Rollingmean_t = \frac{1}{n}\sum_{i=0}^{n-1} PM_{10_{t-i}} \quad (1)$$

$$Rollingstd_t = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}\left(PM_{10_{t-i}} - Rollingmean_t\right)^2} \quad (2)$$

Where n = 5 (rolling window size).

By including these rolling statistics, the model learns from recent patterns and short-term fluctuations, improving its ability to predict sudden PM_{10} peaks.

3.2.5. Interaction terms (Temperature × Humidity, Wind Speed × Pressure)

PM_{10} concentrations are influenced by multiple environmental factors, and their effects are often non-linear. Interaction terms help capture these complex dependencies. This technique was applied in this study due to temperature and humidity interact to affect dust dispersion and particle settling rates. Furthermore, wind speed and pressure influence dust resuspension and transport across the mining site.

The interaction features added to the dataset in this study including:

temp_humidity = temperature × humidity: Aiming to measure the combined effect of temperature and humidity on PM_{10} levels.

wind_pressure = windSpeed × pressure: Aiming to capture the impact of wind force and atmospheric conditions on dust transport.

These interaction terms provide additional information to the model, enabling it to learn more meaningful relationships between environmental factors and PM_{10} concentrations.

3.2.6. Compute and visualize correlation matrix

Computing and visualizing the correlation matrix is a crucial step in this study for removing redundancy (avoids unnecessary computation), enhances model accuracy (by focusing on the most relevant features), prevents overfitting (ensures better generalization to new data), and improves interpretability (helps understand how environmental factors influence PM_{10}). Figure 6 shows the correlation matrix of the PM_{10} dataset used in this study after handling the previous steps.

As shown in Figure 6, we can see that the variables 'temperature', 'wind_pressure', 'rolling_std', and 'pressure' should be removed due to highly correlated features (> 0.8).

3.3. Dimensionality reduction with PCA

Feature engineering introduces a large number of new variables, including lag features, rolling statistics, and interaction terms. While these features enhance the predictive power of machine learning models, they can also introduce multicollinearity and increase computational complexity. To address this issue, Principal Component Analysis (PCA) is applied to reduce dimensionality while retaining 90% of the total variance.

This step is necessary because multicollinearity occurs when two or more features are highly correlated, leading to redundancy in the dataset. Also, highly correlated features can cause overfitting, making the model overly dependent on specific features rather than learning general patterns. In addition, reducing correlation improves model interpretability by keeping only the most relevant features.

To detect highly correlated features, the Pearson correlation coefficient is computed between all variables using Eq. (3):

$$\rho(X,Y) = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2}\cdot\sqrt{\sum(Y_i-\bar{Y})^2}} \quad (3)$$

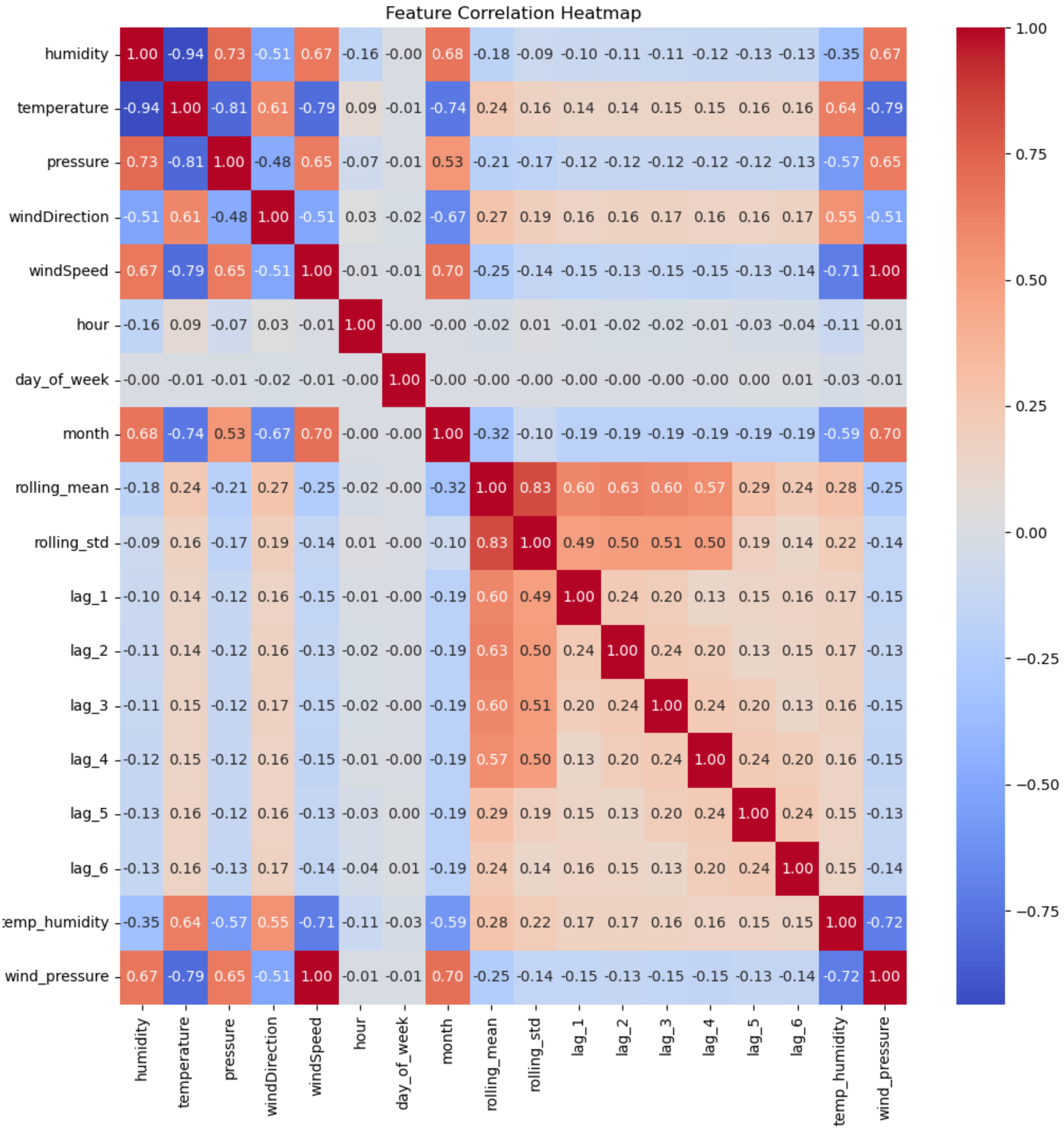


Figure 6. Correlation matrix of the PM_{10} dataset.

Where $\rho(X,Y)$measures the strength of correlation between two features. A value greater than 0.8 indicates high correlation.

To remove highly correlated features, compute a correlation matrix to measure relationships between all features. Then identify pairs of features with correlation > 0.8. Finally, keep only one feature from each highly correlated pair to reduce redundancy.

Subsequently, PCA was applied to reduce dimensionality while retaining 90% variance. PCA is a widely used technique for reducing high-dimensional datasets while preserving the most important patterns in the data. By transforming the original features into a smaller set of uncorrelated principal components, PCA allows the model to learn effectively with fewer variables.

In forecasting PM_{10} in open-pit mines, PCA retains the most significant features while reducing dimensionality, eliminates redundancy from highly correlated variables, enhances computational efficiency for training machine learning models, and improves model generalization by reducing overfitting.

PCA works by transforming the original dataset into a new coordinate system where the first principal component (PC1) captures the most variance, followed by PC2, PC3, etc.

Mathematical representation of PCA is as follows:

1. Compute the covariance matrix using Eq. (4):

$$C = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T \qquad (4)$$

Where X is the dataset and C is the covariance matrix.

2. Compute eigenvalues and eigenvectors using Eq. (5):

$$Cv = \lambda v \qquad (5)$$

Where v are eigenvectors (principal components) and λ are eigenvalues (variance captured by each component).

3. Select top k components that retain 90% variance using Eq. (6):

$$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{m}\lambda_i} \geq 0.90 \qquad (6)$$

Where k is the number of principal components retained.

After computing eigenvalues, we determined how many principal components retain 90% of the variance. Finally, the dataset was transformed from n original features into k principal components, where k retains at least 90% variance. The results showed that the variance retained by PCA is 0.9997, and the columns in the training dataset include 'PC1', 'PC2', 'PC3', 'PC4', 'PC5', 'PC6', 'PC7', 'PC8', 'PC9', 'PC10', and 'dust10' (PM_{10}).

3.4. Machine learning and time-series models

3.4.1. Proposing the RF-PM_{10}Hybrid model

To improve PM_{10} forecasting accuracy in open-pit mines, we propose RF-PM_{10}Hybrid, a hybrid model based on Random Forest (RF), enhanced by feature engineering, SMOGN oversampling, and PCA-based dimensionality reduction. The motivation for this hybrid approach is to leverage the robustness of Random Forest in handling noisy environmental data while addressing common challenges in air quality forecasting, such as class imbalance and temporal dependencies.

The key components of RF-PM_{10}Hybrid include:

- Random Forest (RF): A robust ensemble learning method that averages multiple decision trees to reduce overfitting and improve generalization.

- Feature engineering:

+ Temporal features: Hour of the day, day of the week, and month to capture seasonal trends.

+ Lag features: Historical PM_{10} values as predictors for future concentrations.

+ Rolling statistics: Mean and standard deviation of PM_{10} over a 5-hour window to capture short-term fluctuations.

+ Interaction features: Multiplication of temperature and humidity, as well as wind speed and pressure, to incorporate meteorological dependencies.

- Dimensionality reduction (PCA): Removing highly correlated features (correlation > 0.8) and applying PCA to retain 90% of variance while reducing feature complexity.

- Handling imbalanced data (SMOGN): Since PM_{10} concentration varies significantly, minority extreme PM_{10} values (high and very high levels) are underrepresented. We employ Synthetic Minority Oversampling via Gaussian Noise (SMOGN) to create synthetic data points in these underrepresented regions, leading to a more balanced dataset.

The RF-PM_{10}Hybrid model effectively integrates data-driven learning, feature transformation, and statistical balancing techniques, making it a robust alternative to traditional and boosting-based methods. The workflow of the RF-PM_{10}Hybrid model is proposed in Figure 7.

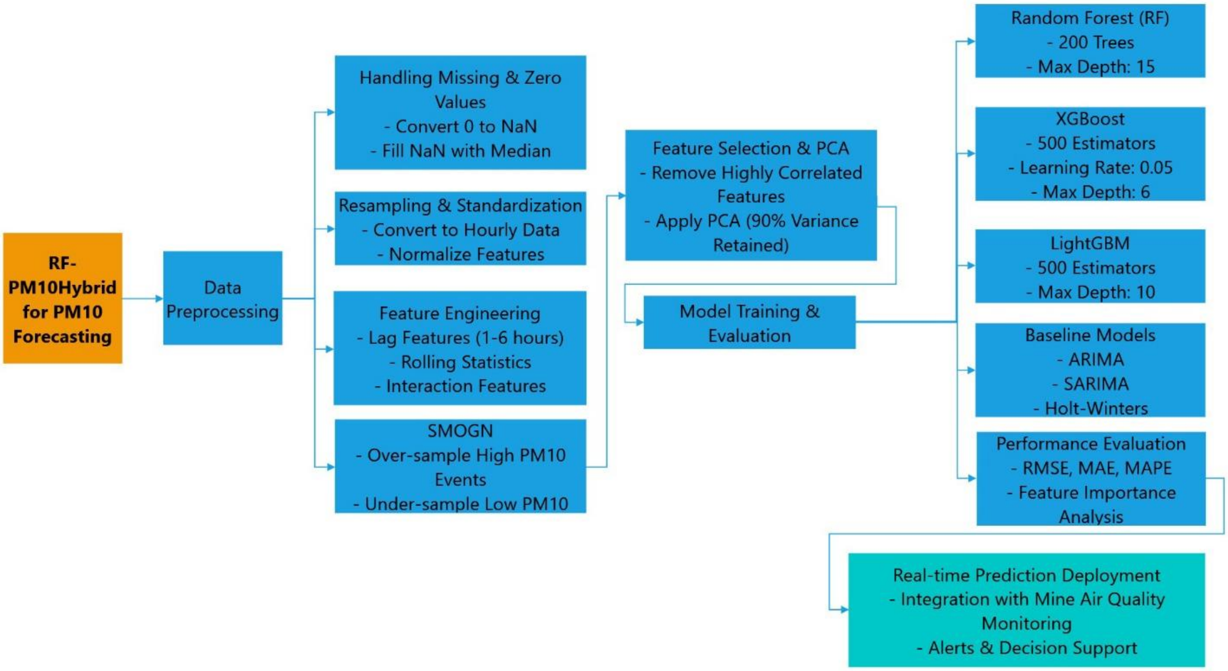

RF-PM10Hybrid for PM10 Forecasting
Data Preprocessing
Handling Missing & Zero Values - Convert 0 to NaN - Fill NaN with Median
Resampling & Standardization - Convert to Hourly Data - Normalize Features
Feature Engineering - Lag Features (1-6 hours) - Rolling Statistics - Interaction Features
SMOGN - Over-sample High PM10 Events - Under-sample Low PM10
Feature Selection & PCA - Remove Highly Correlated Features - Apply PCA (90% Variance Retained)
Model Training & Evaluation
Random Forest (RF) - 200 Trees - Max Depth: 15
XGBoost - 500 Estimators - Learning Rate: 0.05 - Max Depth: 6
LightGBM - 500 Estimators - Max Depth: 10
Baseline Models - ARIMA - SARIMA - Holt-Winters
Performance Evaluation - RMSE, MAE, MAPE - Feature Importance Analysis
Real-time Prediction Deployment - Integration with Mine Air Quality Monitoring - Alerts & Decision Support

Figure 7. Workflow of RF-PM$_{10}$Hybrid for forecasting PM$_{10}$ in this study.

3.4.2. Comparison with other forecasting models

To validate the effectiveness of RF-PM$_{10}$Hybrid, we compare its forecasting performance against several widely used models:

Boosting-based machine learning models

XGBoost: An extreme gradient boosting algorithm that optimizes tree-based learning through regularization and boosting.

LightGBM: A fast, high-performance gradient boosting framework designed for large datasets.

Traditional time-series models

Autoregressive integrated moving average (ARIMA): A parametric time-series forecasting model that captures trends and seasonality.

Seasonal autoregressive integrated moving average (SARIMA): An extension of ARIMA that explicitly models seasonal dependencies.

Statistical smoothing model

Holt-winters exponential smoothing: A forecasting method that applies exponential weights to past observations to model trends and seasonality.

3.5. Performance evaluation metrics

To assess the accuracy and reliability of the PM$_{10}$ forecasting models, we utilize three widely accepted performance evaluation metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These metrics provide insights into different aspects of model performance, including absolute error, percentage error, and error sensitivity to large deviations.

RMSE is a widely used metric for regression problems, measuring the standard deviation of the residuals (prediction errors). It quantifies how much the predicted values deviate from the actual values in terms of squared differences. The formula for RMSE is:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (7)$$

where, y_i, $\hat{y}_i$ and $\bar{y}_i$ denote the values of PM$_{10}$ for measured, predicted, and the mean of the measured values.

MAE measures the average absolute differences between actual and predicted values. Unlike RMSE, it treats all errors equally without squaring them. The formula for MAE is:

$$\mathrm{MAE} = \frac{1}{n}\sum_{\mathrm{i}=1}^{n}|y_i - \hat{y}_i| \qquad (8)$$

MAPE measures the prediction error as a percentage of actual values, making it scale-independent and useful for comparing across different datasets. The formula for MAPE is:

$$\text{MAPE} = \frac{100\%}{n}\sum_{\text{i}=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad (9)$$

4. Results and discussion

4.1. Handling imbalanced data with SMOGN

One of the key challenges in PM_{10} forecasting in open-pit mining is the imbalanced distribution of PM_{10} concentrations. The dataset often contains:

- Long periods of low PM_{10} concentrations, making it difficult for the model to learn about extreme pollution events.

- Short but critical high PM_{10} peaks, typically occurring after blasting, excavation, and material transport activities.

- Underrepresentation of extreme values, leading machine learning models to favor predicting low PM_{10} levels while underestimating high pollution episodes.

Standard machine learning models tend to learn patterns from majority class (low PM_{10} levels) and fail to accurately predict rare but high-impact pollution spikes. Furthermore, underestimating PM_{10} peaks can lead to incorrect air quality warnings, affecting mine safety and regulatory compliance. These led to imbalanced PM_{10} data is a problem in this study.

Normally, Synthetic Minority Over-sampling Technique (SMOTE) is often used for this task. However, SMOTE is designed for classification problems, where the target variable is categorical (e.g., "low PM_{10}" vs. "high PM_{10}"). In this study, PM_{10} dataset is a time-series dataset, and PM_{10} is a continuous variable, meaning we are dealing with a regression problem, not classification. In addition, SMOTE generates synthetic samples by interpolating between existing minority-class points, but it is not optimized for continuous variables with complex distributions. Besides, applying SMOTE to regression problems may create unrealistic PM_{10} values that do not align with physical and environmental constraints.

To overcome these limitations, Synthetic Minority Over-sampling for Regression (SMOGN) was used. Unlike SMOTE, SMOGN is designed for regression problems, allowing it to generate synthetic PM_{10} values while preserving the underlying distribution of the data.

SMOGN works by generating synthetic PM_{10} values in underrepresented regions of the dataset, ensuring that the model can learn to predict both common (low PM_{10}) and rare (high PM_{10}) events. To do this, SMOGN was conducted through the following steps:

Step 1: Identify imbalanced regions

- The dataset is analyzed to determine which PM_{10} values are underrepresented.

- Typically, extreme high PM_{10} values (top 5-10% of the dataset) are underrepresented.

Step 2: Compute distance weights

- SMOGN assigns a higher weight to rare PM_{10} values to ensure that they are oversampled.

- Distance metrics (e.g., Euclidean distance) are used to identify similar samples.

Step 3: Generate synthetic PM_{10} values

- New PM_{10} values are generated by interpolating between real observations in the minority range (high PM_{10} levels).

- This process creates realistic synthetic PM_{10} values that follow the distribution of actual measurements.

Step 4: Combine synthetic and original data

- The new synthetic PM_{10} values are merged with the original dataset, creating a balanced distribution.

To illustrate the effect of SMOGN, we compare PM_{10} distributions before and after applying SMOGN, as shown in Table 3.

Table 3. PM_{10} distribution before and after SMOGN.

PM_{10} range (μg/m^3)	Before SMOGN (Sample count)	After SMOGN (Sample count)
Low (0-50)	2423	875
Moderate (50-100)	630	1102
High (100-150)	183	421
Very high (> 150)	94	211

Observations from the data balancing process in Table 3 showed that the largest reduction occurred in the low PM_{10} range, where the sample count dropped significantly from 2423 to 875. This suggests that SMOGN has performed under-sampling to balance the dataset, as low PM_{10} values were initially dominant. This also prevents the model from being biased toward predicting lower PM_{10} levels, improving its ability to forecast higher values.

The moderate PM_{10} range (50÷100 μg/m^3) increased from 630 to 1102 samples, meaning SMOGN generated synthetic samples to make this category more comparable in size to others. The high PM_{10} range (100÷150 μg/m^3) increased from 183 to 421, nearly doubling in size. The very high PM10 range (>150 μg/m^3) grew from 94 to 211, ensuring that extreme pollution events are adequately represented in the dataset.

Before using SMOGN, the dataset was highly imbalanced, with low PM_{10} values (0÷50 μg/m^3) dominating, while very high values (>150 μg/m^3) were underrepresented. A machine learning model trained on this dataset would likely struggle to predict high PM_{10} values accurately because of the lack of sufficient high PM_{10} samples.

After using SMOGN, the dataset is now better balanced, ensuring the model learns to predict PM_{10} levels across a broader range, not just in low concentrations. In addition, this technique also improved forecasting accuracy, particularly for higher PM_{10} events that are critical for air quality management in open-pit mines, as shown in Figure 8.

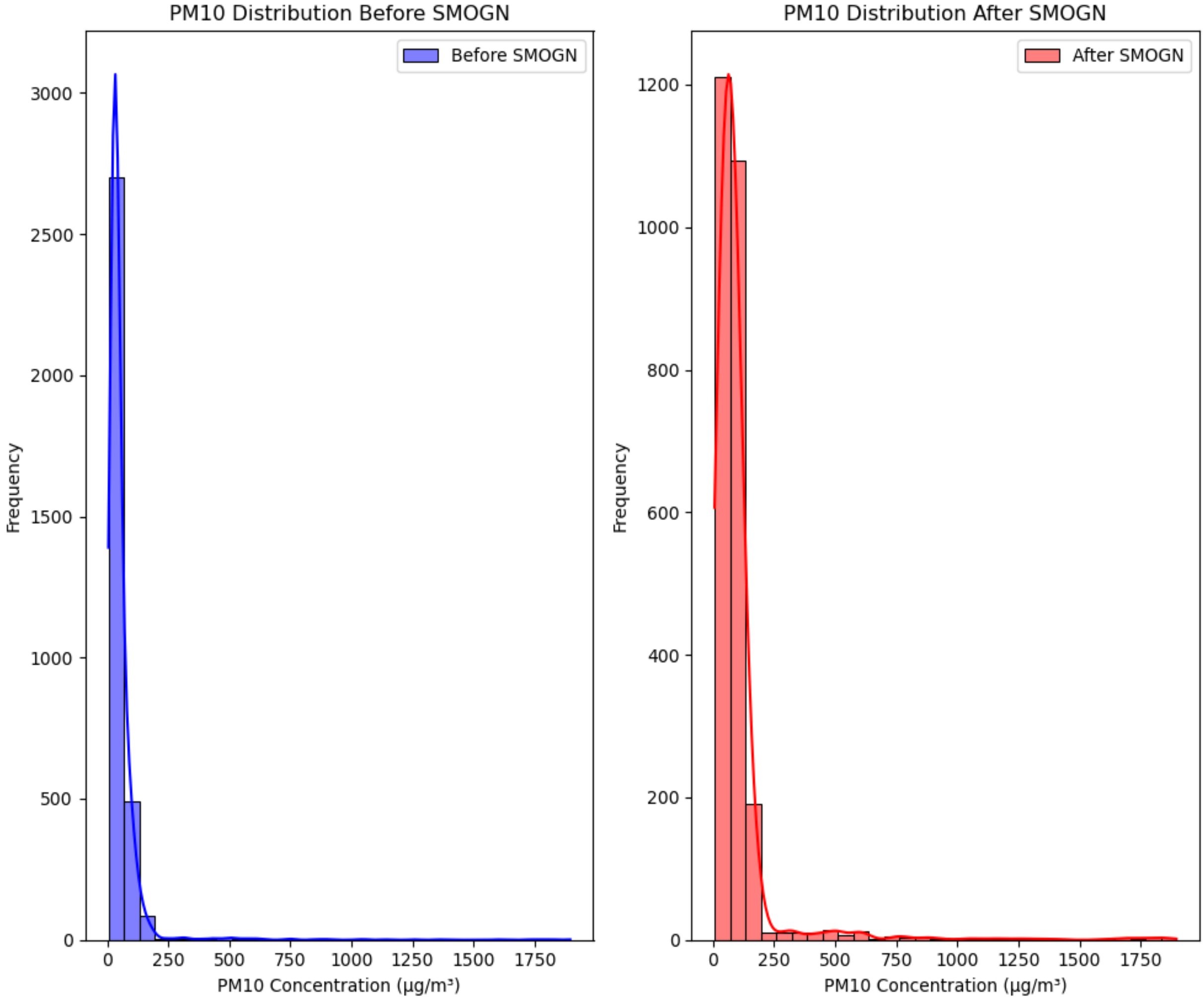


Figure 8. Comparison between before and after SMOGN in this study.

4.2. Feature importance analysis using Random Forest

Once the dataset was balanced, the Random Forest was applied to analyze feature important, aiming to reduce the number of variables used, getting simpler and better model for forecasting PM_{10} at the Sin Quyen copper mine. The results are shown in Figure 9. The results indicated that the following low importance features should be removed: 'PC1', 'PC2', 'PC3', 'PC4', 'PC5', 'PC6', 'PC7', 'PC8', 'PC10'. Finally, only PC9 was used to forecasting PM_{10} in this study.

4.3. Development of the forecast models

This study evaluates three machine learning models and three statistical models for PM_{10} forecasting. For Machine Learning Models, Random Forest (RF), XGBoost (Extreme Gradient Boosting), and LightGBM (Light Gradient Boosting Machine) were developed. Meanwhile, ARIMA (AutoRegressive Integrated Moving Average), SARIMA (Seasonal ARIMA), and Holt-Winters Exponential Smoothing are the statistical models used in this study for comparison.

To do this, five input variables were selected to forecast PM_{10}, including wind speed, wind direction, humidity, temperature, atmospheric pressure. The selection of five meteorological input variables—humidity, temperature, pressure, wind direction, and wind speed—for PM_{10} forecasting in this study is based on their well-established influence on particulate matter behavior in open-pit mining environments. These variables were chosen after an extensive review of previous literature and domain-specific knowledge of dust dynamics in mining operations. Specifically:

- Wind speed and wind direction directly control the dispersion and transport pathways of airborne PM_{10} particles. High wind speeds can resuspend settled dust and carry pollutants over long distances, while wind direction determines the trajectory of dust plumes.

- Humidity affects particle agglomeration and settling. Low humidity typically enhances dust suspension, while high humidity promotes coagulation and deposition of particles.

- Temperature influences atmospheric stability and vertical mixing of air layers. It also interacts with humidity to determine the likelihood of dust re-entrainment.

- Atmospheric pressure is related to air density and influences the vertical movement of dust. Sudden drops in pressure may correlate with increased turbulence and dust emissions.

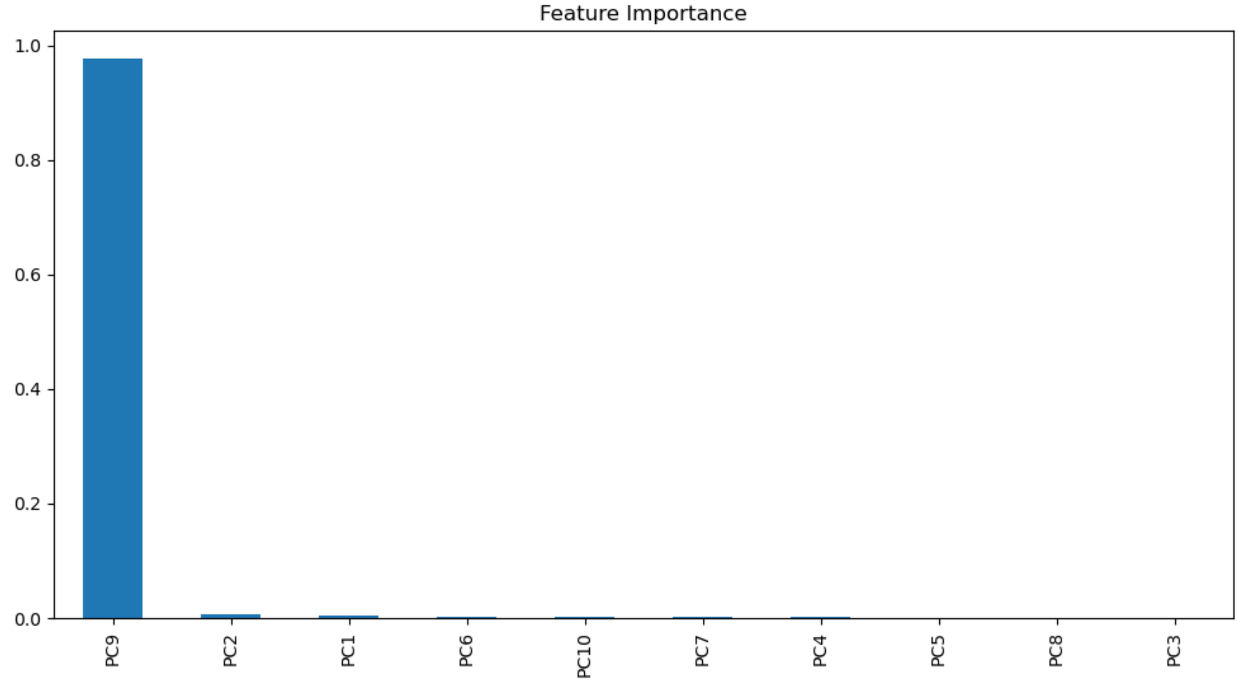


Figure 9. Feature important analysis of the dataset used.

These five variables were not only chosen for their scientific relevance but also because they are continuously monitored at the mine site and readily available in real-time, making them ideal for integration into a real-time forecasting system. In addition, prior studies have demonstrated that these meteorological factors are among the most important predictors in air quality forecasting models using tree-based machine learning methods such as RF, XGBoost, and LightGBM.

Finally, feature importance analysis conducted as part of this study confirmed the predictive relevance of these variables, further justifying their inclusion in the final forecasting models.

For developing the RF-PM$_{10}$Hybrid model, an ensemble model using multiple decision trees. This model handles non-linearity well and is robust to noisy data. The RF-PM$_{10}$Hybrid model was trained with 200 trees, max depth = 15.

For the XGBoost, a boosting algorithm that builds trees sequentially, which was trained with 500 estimators, learning rate = 0.05 and max depth = 6. Early stopping was used to prevent overfitting.

For the LightGBM model, it was optimized for speed and efficiency. It handles large datasets better than traditional boosting models, and was configured with 500 estimators, max depth = 10.

For the development of the ARIMA model, it was trained with order (2,1,2) (auto-selected using AIC criterion).

SARIMA, an extension of ARIMA that accounts for seasonality. The seasonal order was set to (1,1,1,24) to capture daily patterns in this study.

Holt-Winters Exponential Smoothing can capture trend and seasonality in time-series data. It was configured with additive trend and seasonal components.

Before training the models, the dataset was split into two parts with 80% of the whole dataset was used to train the models, and the remaining time stamps (20%) were used for testing the performance of the trained models.

RMSE, MAE and MAPE were calculated to evaluate the models' performance, as shown in Table 4.

The machine learning models—RF-PM$_{10}$Hybrid, XGBoost, and LightGBM—demonstrated superior performance compared to traditional time-series models. RF-PM10Hybrid achieved the lowest RMSE (5.791) and MAE (3.518) on the testing dataset, making it the most accurate model for predicting PM$_{10}$ fluctuations. LightGBM also performed well, with RMSE of 6.172 and MAE of 3.770, confirming the effectiveness of boosting techniques in handling complex air quality datasets.

Among the machine learning models, XGBoost had a slightly higher RMSE (8.293) on the testing dataset, indicating slightly lower generalization performance compared to RF-PM10Hybrid and LightGBM. However, XGBoost maintained a relatively stable RMSE and MAE between the training and testing datasets, suggesting less overfitting compared to LightGBM, which showed a more substantial RMSE gap between training (41.985) and testing (6.172).

RF-PM$_{10}$Hybrid not only outperformed all other models on the testing dataset but also maintained stability between training and testing performance, reducing the risk of overfitting. The MAPE of RF-PM10Hybrid (11.70%) further confirms its reliability in predicting air quality

Table 4. Performance of the developed models for forecasting PM$_{10}$ at the Sin Quyen open-pit copper mine.

Model	Training dataset			Testing dataset		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
RF-PM$_{10}$Hybrid	11.981	7.329	10.96%	5.791	3.518	11.70%
XGBoost	24.032	12.935	18.42%	8.293	3.953	13.18%
LightGBM	41.985	15.367	18.60%	6.172	3.770	12.57%
ARIMA	129.780	37.597	57.27%	4.233	4.208	14.03%
SARIMA	131.516	39.433	65.69%	11.070	8.800	29.32%
Holt-Winters	131.077	41.752	68.05%	13.108	10.224	34.07%

variations in the open-pit mining environment. The success of RF-PM_{10}Hybrid can be attributed to its integration of feature engineering techniques (lag features, rolling statistics, and interaction terms), PCA for dimensionality reduction, and SMOGN for handling data imbalance.

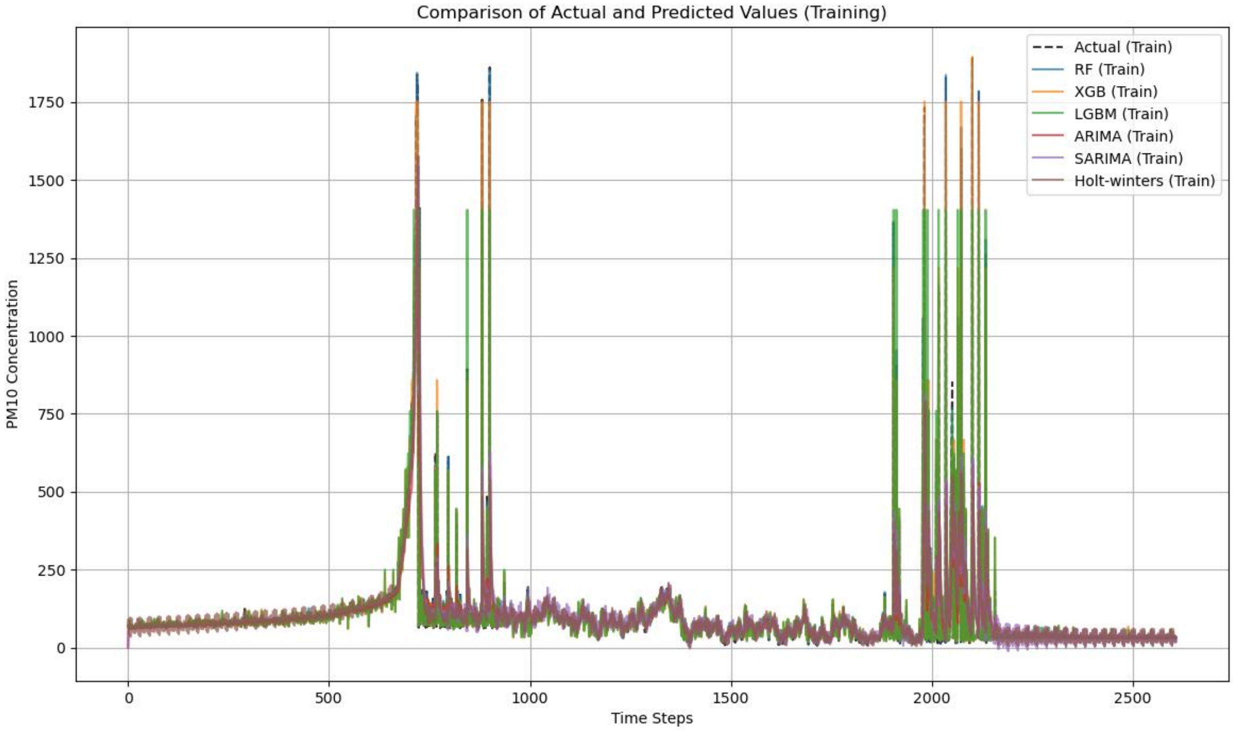
The statistical models—ARIMA, SARIMA, and Holt-Winters—performed significantly worse than machine learning models. ARIMA and SARIMA, in particular, exhibited high RMSE values in the training dataset (129.780 and 131.516, respectively), indicating poor model fit to the PM_{10} data. Additionally, SARIMA produced the highest MAPE (29.32%) on the testing dataset, reflecting its difficulty in handling PM_{10} variability in open-pit mines. Holt-Winters performed the worst overall, with a MAPE of 34.07%, making it unsuitable for accurate forecasting in this context.

The poor performance of ARIMA, SARIMA, and Holt-Winters can be attributed to the high variability and irregularities in PM_{10} data, which traditional time-series models struggle to capture. These models assume stationarity and rely heavily on historical trends, whereas machine learning models can adapt to non-linear relationships and incorporate external meteorological variables such as humidity, wind speed, and temperature.

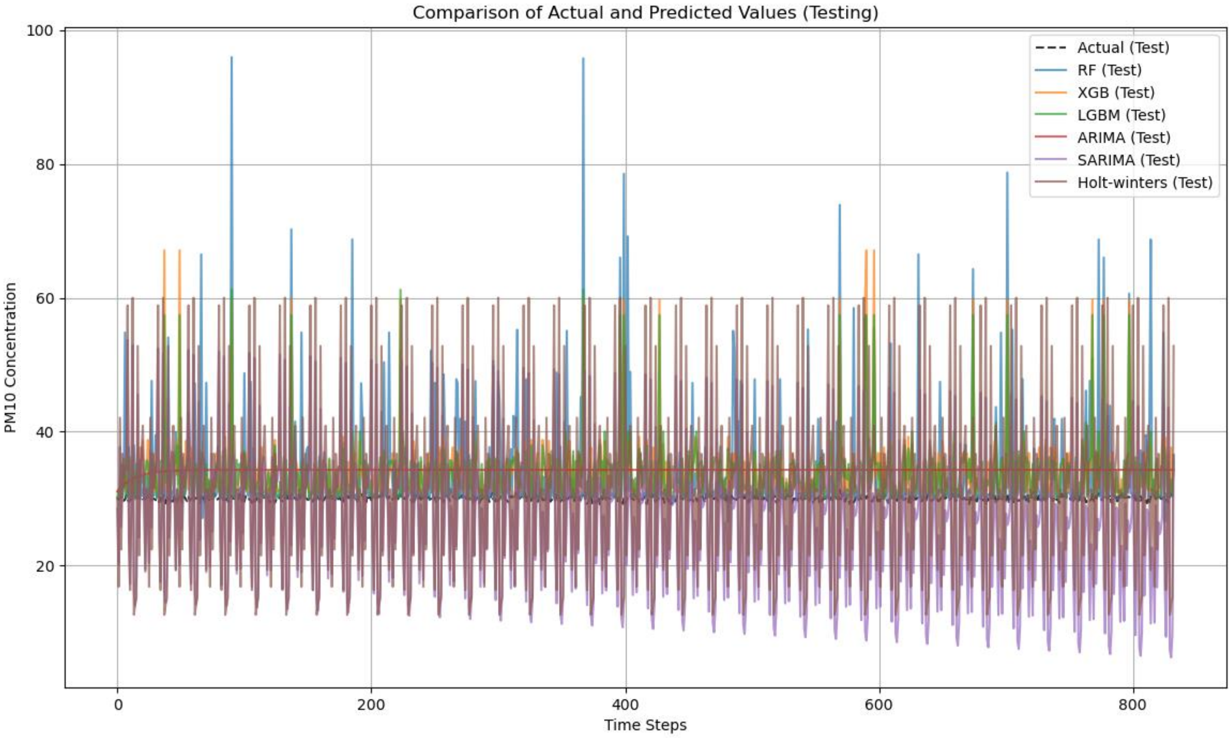
Moreover, the integration of feature engineering techniques (lag features, rolling statistics, and interaction terms) in machine learning models significantly improved their forecasting ability. Statistical models, which rely solely on past PM_{10} values without additional feature inputs, were unable to match the predictive power of ML-based approaches.

One of the key reasons for the improved accuracy of machine learning models was the application of SMOGN (Synthetic Minority Over-sampling Technique for Regression) to handle imbalanced PM_{10} values. Before SMOGN, the dataset was heavily skewed toward lower PM_{10} concentrations, making it difficult for models to accurately predict high PM_{10} events. After applying SMOGN, the distribution of PM_{10} was more balanced, ensuring that the model could better learn extreme pollution conditions caused by mining activities.

To visualize the accuracy of the models developed, Figure 11 shows the comparison of actual and forecasted PM_{10} on both training and testing datasets.



(a)



(b)

Figure 11. Comparison of actual and forecasted PM_{10} on the training and testing datasets, (a) Training dataset; (b) Testing dataset.

The results demonstrate that RF-PM_{10}Hybrid is the most effective model for forecasting PM_{10} in open-pit mining environments, achieving the lowest RMSE, MAE, and MAPE on the testing dataset. The boosting models (XGBoost, LightGBM) also outperformed traditional time-series approaches, confirming the advantages of machine learning over statistical forecasting techniques.

These findings suggest that integrating advanced feature engineering, PCA for dimensionality reduction, and SMOGN for data balancing significantly enhances forecasting accuracy, making RF-PM_{10}Hybrid a valuable tool for real-time air quality management and environmental risk mitigation in open-pit mines.

5. Conclusion and future work

This study developed and evaluated machine learning and statistical models for forecasting PM_{10} concentrations at the Sin Quyen open-pit copper mine, utilizing advanced feature engineering, Principal Component Analysis (PCA), and Synthetic Minority Over-sampling technique for regression (SMOGN) to improve prediction accuracy. The results demonstrated that machine learning models significantly outperformed traditional time-series models in terms of RMSE, MAE, and MAPE. Among the tested models, RF-PM_{10}Hybrid achieved the best overall performance, demonstrating strong predictive capability with the lowest RMSE (5.791) and MAE (3.518) on the testing dataset, followed closely by LightGBM and XGBoost. Conversely, traditional statistical models (ARIMA, SARIMA, and Holt-Winters) struggled to capture the complex variability of PM_{10} concentrations, showing poor generalization and higher forecasting errors. The new findings of this study indicated that machine learning models (RF-PM_{10}Hybrid, XGBoost, LightGBM) significantly outperformed ARIMA, SARIMA, and Holt-Winters in PM_{10} forecasting. RF-PM_{10}Hybrid emerged as the best model due to its ability to handle non-linearity, feature interactions, and high-dimensional meteorological data. Feature engineering, including lag features, rolling statistics, and interaction terms, enhanced the model's predictive accuracy. PCA improved computational efficiency by reducing dimensionality while retaining over 90% of data variance. SMOGN effectively balanced the dataset, improving model performance in forecasting high PM_{10} levels associated with mining activities. Boosting models (XGBoost, LightGBM) generalized well, demonstrating strong forecasting capabilities across both training and testing datasets. These findings highlight the potential of machine learning-based approaches for real-time PM_{10} monitoring, providing an effective decision-support tool for mine operators, environmental policymakers, and regulatory agencies.

While this study has demonstrated promising results in PM_{10} forecasting for open-pit mines, there are several areas for future research and improvement:

- Future models should consider real-time meteorological data (e.g., humidity, wind patterns, temperature fluctuations) with finer resolution to improve predictive accuracy.
- Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Transformer-based architectures can be explored to improve long-term forecasting.
- Future research can integrate traditional dispersion models (e.g., AERMOD) with machine learning approaches to enhance physical interpretability.
- Hybrid models combining statistical methods (e.g., SARIMA) with machine learning (e.g., RF-PM_{10}Hybrid) may further improve forecasting performance.

Acknowledgement

This paper was supported by the Vietnam National Coal and Mineral Industries Holding Corporation Limited (VINACOMIN) in Viet Nam under grant number KC.05.Đ11-23/21-25.

Contributions of authors

Ngoc Tuan Le - methodology, data collection, formal analysis, model development, visualization, validation, writing original draft, writing review & editing; Hoang Nguyen - conceptualization; methodology, data collection, formal analysis, model development, visualization, validation, writing original draft, writing review & editing; Nam Xuan Bui - data collection, formal analysis, visualization, writing original draft, writing review & editing; Hoa Thu Thi Le - data collection, formal analysis, visualization.

References

Asselman, A., et al. (2023). "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm." Interactive Learning Environments 31(6): 3360-3379.

Bhatti, U. A., et al. (2021). "Time series analysis and forecasting of air pollution particulate matter (PM 2.5): an SARIMA and factor analysis approach." IEEE Access 9: 41019-41031.

da Silva, K. L. S., et al. (2023). "Spatio-temporal visualization and forecasting of PM 10 in the Brazilian state of Minas Gerais." Scientific reports 13(1): 3269.

Fratello, M. and R. Tagliaferri (2018). "Decision trees and random forests." Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics 1(S3): 374.

Halabaku, E. and E. Bytyçi (2024). "Overfitting in Machine Learning: A Comparative Analysis of Decision Trees and Random Forests." Intelligent Automation & Soft Computing 39(6).

Hegelich, S. (2016). "Decision trees and random forests: Machine learning techniques to classify rare events." European policy analysis 2(1): 98-120.

Kavitha, R. and M. Priyadharshini (2024). Performance Comparison of XGBoost and LightGBM Gradient Boosting Algorithms in Predicting Cervical Cancer Risk. 2024 International Conference on Computing and Data Science (ICCDS), IEEE.

Bui, X, N, (2021). Development of air quality control system to ensure safety and healthy in deep open-pit mine in Quang Ninh area (in Vietnamese).

Pozza, S. A., et al. (2010). "Time series analysis of PM2. 5 and PM10− 2.5 mass concentration in the city of Sao Carlos, Brazil." International Journal of Environment and Pollution 41(1-2): 90-108.

Sánchez Lasheras, F., et al. (2020). "Evolution and forecasting of PM10 concentration at the Port of Gijon (Spain)." Scientific reports 10(1): 11716.

Sibindi, R., et al. (2023). "A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices." Engineering Reports 5(4): e12599.

Simon, S. M., et al. (2023). "Interpreting random forest analysis of ecological models to move from prediction to explanation." Scientific reports 13(1): 3881.

Sumanth, C., et al. (2020). "“Numerical modelling of PM10 dispersion in open-pit mines”." Chemosphere 259: 127454.

Török, Z., et al. (2023). "Modelling the dispersion of particulate matter (PM10) via wind erosion from opencast mining—Moldova Nouă tailings ponds, Romania." Environmental monitoring and assessment 196(1): 59.

Wang, F., et al. (2021). "Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation." Environmental research 202: 111660.

Wang, Y., et al. (2025). "An interpretable approach combining Shapley additive explanations and LightGBM based on data augmentation for improving wheat yield estimates." Computers and Electronics in Agriculture 229: 109758.

Zhang, D. and Y. Gong (2020). "The comparison of LightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure." IEEE Access 8: 220990-221003.