Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Modeling Dust Emissions Using ANN, XGBoost, and Random Forest Techniques: A Case Study of the Meftah Cement Plant (Algeria)

B. Touahar, Y. Kerchich, R. Kerbachi, Y. Medkour, M.A. Bouda, A. Teffahi[1], A. Djouahi[2], I. Kemerchou[2]

[1]Laboratory of Environmental Science and Technology

Department of Environmental Process Engineering, Ecole Nationale Polytechnique, Algiers, Algeria

[2] Kasdi Merbah Ouargla University, BP 511, Ouargla 30000, Algeria

Corresponding Author: bachir.touahar@univ-ouargla.dz (B. TOUAHAR).

ARTICLE INFO

Received: 01 Dec 2025

Revised: 09 Jan 2026

Accepted: 20 Jan 2026

ABSTRACT

The production of clinker in cement plants remains a significant source of atmospheric dust emissions, posing a major challenge for compliance with environmental standards and the protection of public health. This research evaluates and compares the performance of three machine learning architectures—Artificial Neural Networks (ANN), XGBoost, and Random Forest—to predict dust levels from the AFF2 stack at the Meftah cement plant (Blida, Algeria) in near real-time. The predictive models are developed based on five key operational variables: burned gas rate, exhaust gas temperature at the tower outlet, raw meal feed rate (flour), filter differential pressure (AFF2), and excess air percentage (O2). For model training and validation, an exhaustive dataset comprising over 13,000 historical observations from 2023 and 2024 was utilized. The obtained results demonstrated that ensemble tree-based models significantly outperformed the ANN model. Among the investigated approaches, the Random Forest model achieved the best predictive performance with the lowest error values (MSE = 8.854, RMSE = 2.976, and MAE = 1.941) and the highest coefficient of determination (R^2 = 0.750). The XGBoost model also produced strong predictive capability with an R^2 value of 0.736, while the ANN model showed comparatively lower performance with an R^2 value of 0.645. Residual analysis and prediction-versus-actual plots further confirmed the superior robustness and generalization capability of the Random Forest algorithm. The findings demonstrate the effectiveness of machine learning techniques for real-time prediction of dust emissions in cement manufacturing processes. The proposed predictive framework can serve as an intelligent decision-support tool for environmental monitoring, predictive maintenance, and proactive control of particulate emissions. The implementation of such models may help cement plants reduce environmental impacts, optimize filtration system performance, and improve compliance with environmental regulations.

Keywords: Artificial Neural Networks, XGBoost, Random Forest, Cement Plant, Dust Emissions, AFF2 Stack, Modeling.

Introduction

The global cement industry, responsible for approximately 2.4 billion tonnes of CO_2 and significant particulate matter (PM) annually, is undergoing a rapid digital transformation to meet ambitious sustainability targets (Fayaz et al., 2026). In emerging economies like Algeria, this transition is not merely an environmental goal but a strategic economic necessity as the nation seeks to modernize its

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

industrial fleet (Ministry of Industry, 2023; ASJP, 2025). The Meftah cement plant, particularly its AFF2 stack, serves as a critical case study for implementing advanced monitoring solutions that go beyond traditional steady-state models.

Predicting dust emissions is inherently challenging due to the high variability of operational data across different industrial sites, where prediction errors can vary significantly based on data richness (Fayaz et al., 2026). While traditional chemical transport models and empirical equations often fail to capture real-time fluctuations, machine learning (ML) architectures have proven to be much more robust in building accurate 24-h and 48-h pollutant forecasts (MDPI, 2023). However, a significant gap remains in identifying which specific ML architecture—Artificial Neural Networks (ANN), XGBoost, or Random Forest—provides the optimal balance between predictive accuracy and computational efficiency in the specific context of Algerian clinker production.

Recent studies indicate that while ANN models excel at modeling complex non-linearities, ensemble methods like XGBoost and Random Forest often offer superior performance on structured industrial datasets by reducing overfitting and providing better handling of missing data (MDPI, 2025; ResearchGate, 2025). This article addresses this gap by conducting a rigorous benchmarking of these three models using a massive dataset of 13,000 observations from the Meftah plant (2023–2024). By evaluating key performance metrics such as R^2, RMSE, and MAE, this research aims to provide a "surrogate modeling" framework that allows operators to anticipate emission overshoots before they occur, thereby ensuring continuous compliance with Executive Decree No. 06-138 (Algerian regulations) and promoting a data-driven path toward low-emission industrial operation.

State of the Art

1.1. The Shift Toward Data-Driven Industrial Monitoring

Traditional methods for monitoring particulate matter (PM) in cement manufacturing have historically relied on stationary continuous emission monitoring systems (CEMS) and empirical linear models. However, these approaches often struggle to account for the high non-linearity and "data richness" of modern clinker production lines (Fayaz et al., 2026). Recent literature suggests that prediction errors in industrial stacks can vary by up to 3 to 5 times depending on the complexity of the operational parameters used (Fayaz & Sheikh, 2026). Consequently, the industry is shifting toward surrogate modeling, which uses process data (temperature, pressure, gas flow) to predict emissions in near real-time.

1.2. Benchmarking Machine Learning Architectures

Current research (2024–2026) emphasizes the benchmarking of three primary algorithmic families for air quality and industrial emissions:

1. **Artificial Neural Networks (ANN):** Recent studies demonstrate that ANNs remain superior in capturing extreme pollution events and highly non-linear relationships. In comparative studies for PM estimation, ANNs improved performance metrics (such as Nash–Sutcliffe efficiency) by over 40% compared to linear models (MDPI, 2025).

2. **XGBoost (eXtreme Gradient Boosting):** This model has emerged as a state-of-the-art solution for structured industrial data. Recent evaluations in similar industrial contexts have shown XGBoost achieving near-perfect accuracy ($R^2 = 0.99$) due to its ability to handle feature sparsity and prevent overfitting through gradient boosting frameworks (PLOS One, 2025).

3. **Random Forest (RF):** Often cited for its robustness and interpretability, RF is frequently used as a baseline for ensemble learning. While it occasionally underperforms compared to XGBoost in high-dimensional time-series data, it remains a gold standard for its stability across varying operational conditions (ResearchGate, 2025).

1.3. The Algerian Context and Regulatory Imperatives

In Algeria, the application of these advanced models is becoming a regulatory necessity. Research specifically focusing on the Algerian cement sector highlights that ML-based surrogate models can achieve a Mean Absolute Error (MAE) of approximately 2.37 mg/m³, which is well within the safety margins of the Executive Decree No. 06-138 (30–50 mg/Nm³) (ResearchGate, 2025). Despite these

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

advancements, there is limited research comparing these three specific architectures (ANN, XGBoost, and RF) using large-scale, multi-year datasets (e.g., >13,000 observations) specifically for the AFF2 stack at the Meftah plant.

Materials and Methods

1.4. Data Collection and Variable Selection

The study utilizes a high-resolution dataset extracted from the Meftah cement plant's monitoring systems, covering a continuous period through 2023 and 2024. The final dataset comprises 13,052 observations sampled at hourly intervals. To predict dust concentration (mg/Nm³) from the AFF2 stack, five key operational variables were selected based on their physical influence on filter efficiency:

Table 1. keys operational variables.

Variable	Description	Unit
Gas Flow Rate	The volume of burned gas passing through the filter.	m^3/h
Exhaust Temperature	Temperature at the tower outlet/filter inlet.	°C
Kiln Feed (Flour)	The mass flow rate of raw meal introduced into the kiln.	t/h
Differential Pressure (Delta P)	The pressure drop across the AFF2 filter bags.	mbar
Excess Air (O_2)	The percentage of oxygen in the exhaust gases.	%

1.5. Data Preprocessing

Before feeding the data into the machine learning algorithms, a rigorous preprocessing pipeline was implemented:

- **Data Cleaning:** Outliers resulting from sensor malfunctions or plant shutdowns were identified and removed using the Interquartile Range (IQR) method.
- **Normalization:** Since the variables operate on different scales (e.g., temperature vs. gas flow), Min-Max Scaling was applied to transform all features into a range of [0,1], ensuring faster convergence for the ANN.
- **Data Splitting:** The dataset was partitioned into two subsets: 80% for training (to build the models) and 20% for testing (to evaluate generalization performance on unseen data).

1.6. Model Architectures and Hyperparameters

Three distinct algorithms were configured and optimized:

1.6.1. Artificial Neural Network (ANN)

The implemented ANN is a Multi-Layer Perceptron (MLP). For each layer l, the transformation of the input vector $a^{(l-1)}$ into the output $a^{(l)}$ is governed by the following equations:

1. **Linear Combination:**

$$z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)} \qquad \textbf{(01)}$$

Where :

- $z^{(l)}$: *The pre-activation linear output of layer l.*
- $W^{(l)}$ = *Weight matrix of layer l*
- $a^{(l)}$ = *Activation output at layer l*
- $b^{(l)}$ = *Bias vector of layer l*

2. **Batch Normalization:** To stabilize the hidden states, the linear output is normalized:

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

$$\widehat{z^{(l)}} = \gamma^{(l)} \left(\frac{z^{(l)} - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \right) + \beta^{(l)} \qquad \textbf{(02)}$$

Where :

- $\widehat{z^{(l)}}$*: The normalized activation of layer l*
- μ_B*: The mean of the linear outputs for the current mini-batch.*
- σ_B^2*: The variance of the linear outputs for the current mini-batch.*
- ϵ*: A very small constant (e.g.,* 10^{-7}*) added for numerical stability to avoid division by zero.*
- $\gamma^{(l)}$*: The learnable scale parameter (weight) that allows the network to rescale the normalized data.*
- $\beta^{(l)}$*: The learnable shift parameter (bias) that allows the network to shift the normalized data.*

3. **Non-linear Activation:** The **ReLU** function is applied to the normalized output:

$$a^{(l)} = \max(0, \widehat{z^{(l)}}) \qquad \textbf{(03)}$$

Where :

- $a^{(l)}$ *= Activation output at layer l*
- *max (0, ·) : The Rectified Linear Unit (ReLU) operation, which replaces all negative values with zero to introduce non-linearity and mitigate the vanishing gradient problem.*
- $\widehat{z^{(l)}}$*: The normalized activation of layer l*

4. **Dropout:** To prevent overfitting, a Bernoulli mask $m^{(l)} \sim$ Bernoulli(0.8) is applied during training:

$$\widetilde{a^{(l)}} = a^{(l)} \cdot m^{(l)} \qquad \textbf{(04)}$$

Where :

- $\widetilde{a^{(l)}}$*: The "thinned" activation output after dropout has been applied.*
- $a^{(l)}$*: the output of the neurons in layer l immediately after the activation function (ReLU)*
- *m (l): The Bernoulli mask vector.*

5. Final Output Layer

The final output $\hat{y}$ (dust concentration) is produced by a linear output layer:

$$\hat{y} = W^{(out)} a^{(final)} + b^{(out)} \qquad \textbf{(05)}$$

Where:

- $\hat{y}$*: Model Output / Predicted Value*
- $W^{(out)}$*: Output Weight Matrix*
- $a^{(final)}$*: Final Hidden Layer Activation*
- $b^{(out)}$*: Output Bias*

1.6.2. Random Forest (RF)

The Random Forest regressor is an Ensemble Bagging model. It generates $B = 300$ independent decision trees. Each tree T_b is trained on a bootstrap sample of the dataset D_b. The final prediction for the AFF2 stack emissions is the arithmetic mean of all individual tree outputs:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x; \Theta_b) \qquad \textbf{(06)}$$

Where:

- $\hat{y}$*: Model Output / Predicted Value*
- B*: Total number of trees (e.g., 500 in this study)*

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- x: *The input vector (gas flow, O_2, temperature, flour rate, ΔP).*
- Θ_b: *The random parameters of the b-th tree, including the subset of features selected at each node split to minimize the Mean Squared Error (MSE).*

1.6.3. XGBoost (Extreme Gradient Boosting)

XGBoost was implemented as a tree-based ensemble method that minimizes a regularized objective function. It builds models sequentially, where each new tree corrects the errors of the previous ones using a gradient descent algorithm. This model is particularly effective for capturing non-linear patterns in structured industrial data without requiring extensive feature engineering.

Unlike the parallel nature of Random Forest, XGBoost is an Additive Model that builds trees sequentially. The prediction at iteration K is the sum of the previous prediction and a new function $f_K(x)$:

$$\widehat{y_i^{(K)}} = \sum_{k=1}^{K} f_k(x_i) = \widehat{y_i^{(K-1)}} + \eta f_K(x_i) \qquad (07)$$

Where $\eta = 0.05$ is the learning rate specified in the code. To find the optimal f_K, the model minimizes a regularized objective function:

$$\mathcal{L}^{(\mathcal{K})} = \sum_{i=1}^{n} L\left(y_i, \widehat{y_i^{(K-1)}} + f_K(x_i)\right) + \Omega(f_K) \qquad (08)$$

The complexity penalty Ω prevents overfitting by penalizing the number of leaves (T) and the magnitude of the weights (w):

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda |w|^2 \qquad (09)$$

Where:

- L: *The loss function (e.g., Mean Squared Error).*
- T: *The number of leaves in the tree.*
- w: *The vector of scores (weights) on the leaves.*
- γ, λ: *Regularization parameters used to prevent overfitting, making XGBoost superior for the noisy sensor data from the Meftah plant.*

1.7. Model Evaluation Metrics

The model is evaluated using four main metrics:

- MSE (Mean Squared Error): Mean squared error

$$\mathbf{MSE} = (1/n)\,\Sigma(\mathbf{yi}\text{-}\hat{\mathbf{y}}\mathbf{i})^2 \qquad (10)$$

- RMSE (Root Mean Squared Error): Square root of the mean squared error

$$\mathbf{RMSE} = \sqrt{(1/n)\Sigma(yi - \hat{y}i)^2} \qquad (11)$$

- MAE (Mean Absolute Error): Mean absolute error

$$\mathbf{MAE} = (1/n)\,\Sigma\,|yi - \hat{y}i| \qquad (12)$$

- R^2 (Coefficient of determination): Measures the quality of the prediction

$$R^2 = 1 - [\Sigma(yi - \hat{y}i)^2 / \Sigma(yi - \bar{y})^2] \qquad (13)$$

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Where:

- y_i = *Observed value for the i-th observation*
- $\hat{y}\_i$ = *Predicted value from the model for the i-th observation*
- $\bar{y}$ = *Mean of observed values,* $\bar{y} = (1/n)\Sigma y_i$
- n = *Total number of observations*

1.8. SHAP (SHapley Additive exPlanations) Analysis

To explain these models, the program uses Shapley values, which distribute the total prediction variance among the five input features. The SHAP value ϕ_j for feature j is calculated as:

$$\boldsymbol{\phi_j} = \sum_{S \subseteq \{x_1,\ldots,x_5\} \setminus \{j\}} \frac{|S|!(5-|S|-1)!}{5!} [f(S \cup \{j\}) - f(S)] \qquad \textbf{(14)}$$

Where:

- φ_i = *SHAP value of feature i*
- F = *set of all input features*
- S = *subset of features not including i*
- $f\_S(x\_S)$ = *model prediction using only features in subset S*
- $|S|$ = *number of features in subset S*

This ensures that the "importance" of parameters like dp filter aff2 or gas flow rate is mathematically grounded in coalitional game theory.

Results and Discussion

1.9. Analysis of Convergence and Learning (Learning Curves)

The learning curves presented in Figure 1 illustrate the evolution of the training loss and validation loss of the Artificial Neural Network (ANN) model during the prediction of dust concentration.

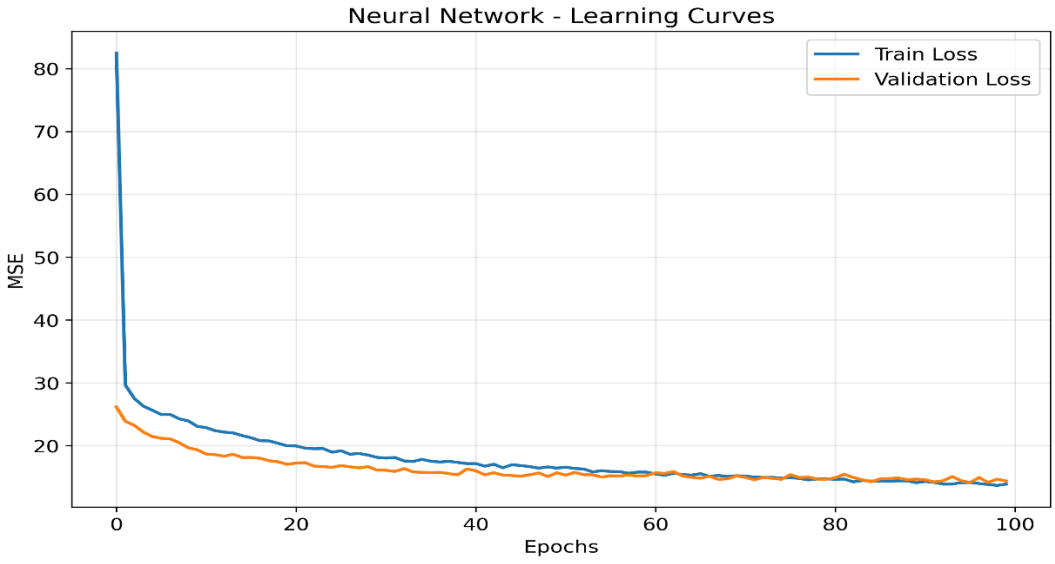

Fig. 1: Evolution of loss functions (MSE) during ANN model training.

At the beginning of the training process, the training loss exhibited a very high value exceeding 80 MSE units, followed by a rapid decrease during the first epochs. This behavior indicates that the neural network quickly learned the dominant nonlinear relationships between the process variables and the dust concentration. Such rapid convergence during early epochs is commonly observed in deep learning models when the optimizer effectively minimizes the prediction error (Goodfellow et al., 2016).

After approximately 20–30 epochs, both the training and validation losses continued to decrease gradually and stabilized around values between 14 and 16 MSE units. The close proximity between the

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

two curves throughout the training process demonstrates that the model generalized well to unseen data and did not suffer from significant overfitting. According to Bishop (2006), a small gap between training and validation errors is generally considered an indicator of good generalization capability in neural network modeling.

Furthermore, the validation loss remained slightly lower than the training loss during several epochs. This phenomenon can occur when regularization techniques such as Dropout and Batch Normalization are applied, since these methods reduce overfitting and improve model robustness during training (Srivastava et al., 2014). The use of Dropout layers likely contributed to preventing excessive memorization of the training dataset while improving predictive stability.

The absence of a significant increase in validation loss at later epochs indicates that the ANN model maintained stable learning behavior throughout the optimization process. Therefore, the selected architecture appears suitable for modeling the complex nonlinear dynamics governing particulate emissions in cement manufacturing processes.

Overall, the learning curves confirm that the ANN model achieved stable convergence, satisfactory learning performance, and good predictive generalization for AFF2 dust concentration estimation.

1.10. Comparative evaluation of predictive performance

Table 2 presents the predictive performance of the Artificial Neural Network (ANN), Random Forest (RF), and XGBoost models developed for estimating dust concentration.

Table. 2: comparative predictive performance.

Model	MSE	RMSE	MAE	R2
ANN	12,5837	3,547351	2,496116	0,644551
Random Forest	8,854234	2,975606	1,940753	0,749896
XGBoost	9,334203	3,055193	2,096569	0,736339

The obtained results indicate that the Random Forest model achieved the best overall predictive performance among the three investigated machine learning approaches. Specifically, the RF model produced the lowest prediction errors with an MSE of 8.854, RMSE of 2.976, and MAE of 1.941, while simultaneously achieving the highest coefficient of determination (R^2 = 0.750). These results suggest that Random Forest was more effective in capturing the nonlinear relationships between the operating parameters and the dust concentration values.

The superior performance of Random Forest can be attributed to its ensemble learning mechanism, which combines multiple decision trees to reduce variance and improve prediction robustness (Breiman, 2001). In industrial process modeling, RF models are widely recognized for their strong generalization capability and resistance to overfitting, particularly when dealing with complex and noisy datasets.

The XGBoost model also demonstrated strong predictive capability, achieving an R^2 value of 0.736 with relatively low prediction errors (RMSE = 3.055 and MAE = 2.097). Although its performance was slightly lower than that of Random Forest, XGBoost remained highly competitive. This behavior is consistent with previous studies showing that gradient boosting algorithms are highly efficient for nonlinear regression problems due to their sequential error-correction mechanism (Chen & Guestrin, 2016).

In contrast, the ANN model exhibited the lowest predictive performance among the three models, with an R^2 value of 0.645 and the highest prediction errors (RMSE = 3.547 and MAE = 2.496). Although the neural network was capable of learning nonlinear process dynamics, its performance may have been influenced by factors such as dataset size, hyperparameter selection, and sensitivity to data variability. Neural networks generally require larger datasets and extensive parameter optimization to outperform ensemble tree-based methods in industrial applications (Goodfellow et al., 2016).

Overall, the comparative analysis demonstrates that ensemble tree-based models, particularly Random Forest, are more suitable for predicting AFF2 dust emissions under the considered operating conditions.

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The results confirm that machine learning techniques can effectively model particulate emissions in cement manufacturing processes and support environmental monitoring and process optimization.

1.11. Correlation Analysis: Observed Values vs. Predicted Values

Figure 2 presents the comparison between the predicted and actual dust concentration values obtained using the Artificial Neural Network (ANN), Random Forest (RF), and XGBoost models for AFF2 stack emission prediction.

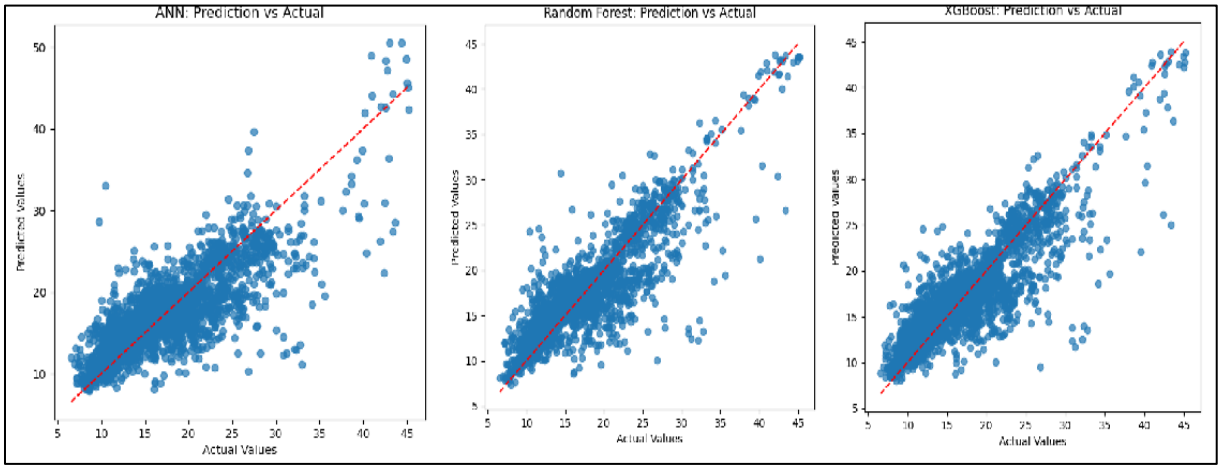

Fig. 2: Predicted values based on the actual values of dust concentration.

In these scatter plots, the red dashed diagonal line represents the ideal prediction line, where predicted values are exactly equal to the actual measured values. Therefore, the closer the data points are to this line, the better the predictive performance of the model (Bishop, 2006).

The ANN model exhibited a relatively wider dispersion of points around the diagonal line, particularly for higher dust concentration values. This indicates that the ANN predictions were less accurate and showed larger deviations from the real measurements. The dispersion suggests that the neural network had greater difficulty capturing the variability of the industrial process under certain operating conditions. This observation is consistent with the lower coefficient of determination obtained for ANN (R^2 = 0.645). According to Goodfellow et al. (2016), neural networks may require large datasets and extensive hyperparameter optimization to achieve robust generalization performance in complex industrial systems.

In contrast, the Random Forest model showed a significantly stronger alignment of data points along the ideal prediction line. Most predictions closely followed the diagonal trend, indicating higher predictive accuracy and better agreement between predicted and measured dust concentrations. The reduced scatter around the line confirms the strong generalization capability of the RF model and its ability to effectively model nonlinear relationships in cement process emissions. Breiman (2001) reported that Random Forest models are highly effective in handling nonlinear and noisy industrial datasets due to their ensemble averaging mechanism. These visual observations are supported by the superior statistical indicators obtained previously, including the highest R^2 value (0.750) and the lowest RMSE and MAE values.

Similarly, the XGBoost model demonstrated good predictive behavior, with most points concentrated near the diagonal line. However, a slightly larger dispersion was observed compared with Random Forest, especially in intermediate and high concentration regions. Although XGBoost successfully captured the global trend of the data, its predictions were marginally less accurate than those of the RF model. This interpretation agrees with the quantitative performance metrics where XGBoost achieved an R^2 value of 0.736. Chen and Guestrin (2016) explained that XGBoost improves prediction accuracy through sequential boosting and gradient optimization, making it particularly efficient for nonlinear regression tasks.

Overall, the graphical analysis confirms that the ensemble tree-based methods (Random Forest and XGBoost) provided better predictive capability than the ANN model for AFF2 dust concentration prediction. Among the evaluated approaches, Random Forest produced the most accurate and stable

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

predictions, making it the most suitable model for monitoring particulate emissions in the studied cement manufacturing process.

1.12. Analysis of the distribution of residues

Figure 3 illustrates the residual distributions obtained for the Artificial Neural Network (ANN), Random Forest (RF), and XGBoost models.

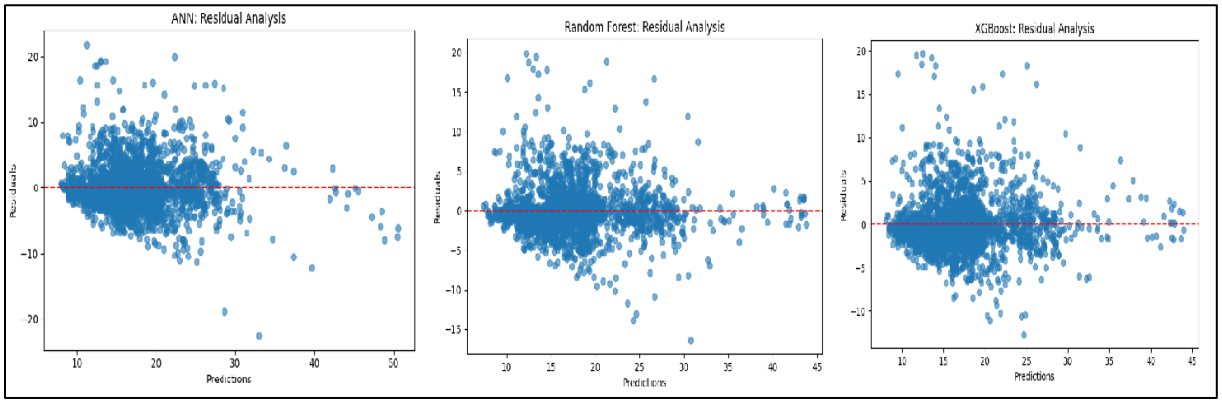

Fig. 3: Residual distributions.

Residual analysis is an important diagnostic tool used to evaluate regression model quality because it measures the difference between observed and predicted values. Ideally, residuals should be randomly distributed around the zero line without systematic trends, indicating unbiased predictions and good model generalization capability (Elhishi et al., 2023).

For the ANN model, the residuals exhibited relatively high dispersion around the zero-error line, with several extreme positive and negative values. The spread of residuals increased particularly for higher prediction values, indicating reduced predictive stability and possible heteroscedastic behavior. This suggests that the ANN model had difficulty fully capturing the complex nonlinear relationships governing particulate emissions in the cement manufacturing process. Similar limitations of ANN models in industrial prediction tasks have been reported in recent machine learning studies when datasets are relatively limited or highly variable (Elhishi et al., 2023).

In contrast, the Random Forest model displayed the most favorable residual distribution among all investigated models. Most residuals were tightly concentrated around zero with relatively low variance, indicating highly accurate and stable predictions. The absence of clear systematic structures confirms that the RF model successfully captured the nonlinear interactions between process variables and dust concentration. Recent studies in cement industry applications demonstrated that Random Forest models provide strong robustness and superior generalization performance for industrial process prediction and emission modeling (Kim et al., 2024).

The XGBoost model also showed satisfactory residual behavior, with most residuals centered near zero. However, compared with Random Forest, the residual dispersion was slightly larger for medium and high predicted concentrations. Although XGBoost effectively modeled the global nonlinear trends of the process, the prediction variability remained marginally higher than that observed for RF. Recent research has shown that XGBoost achieves high predictive capability in engineering and cement-related applications because of its boosting-based sequential learning strategy and gradient optimization mechanism (Safhi et al., 2023; Al-Taai et al., 2023).

Overall, the residual analysis confirms the superiority of ensemble tree-based algorithms over the ANN model for AFF2 dust concentration prediction. Among all tested approaches, Random Forest produced the most stable residual distribution and the lowest prediction variability, confirming its suitability for industrial emission monitoring and predictive environmental control in cement manufacturing systems. These findings are also consistent with recent studies in cement process prediction, where Random Forest frequently achieved higher robustness and predictive accuracy than alternative machine learning models (RFR outperforming XGBoost and other methods) in kiln emission prediction applications.

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

1.13. Interpretability of models and importance of variables (SHAP Analysis)

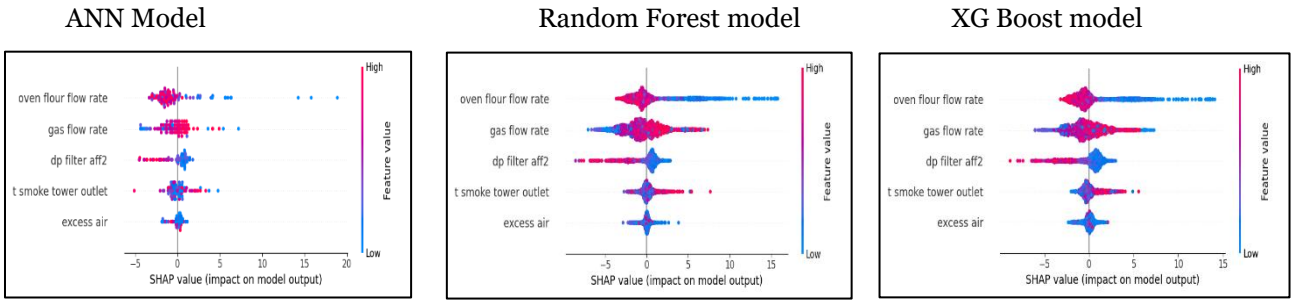

Fig. 4: Importance of input variables based on SHAP values.

Common Dominant Feature: Oven Flour Flow Rate

In all three models, oven flour flow rate consistently appears as the most important feature (highest SHAP value magnitude: 13.0 in XGB, 15.0 in RF, and top position in ANN). This aligns with findings by Ciobanu et al. (2021) and Su et al. (2025), who identified raw meal and fuel flow rates as primary determinants of dust and NOx emissions in cement kilns. The high SHAP values indicate that even small changes in this flow rate can significantly alter model predictions — likely due to its direct influence on combustion efficiency and particulate matter generation.

Gas Flow Rate and Pressure Drops

- Gas flow rate ranks second in XGB (12.0) and RF (11.0), and third in ANN (SHAP = -1.5 in relative ordering).
- dp filter aff2 (likely differential pressure across a baghouse or cyclone filter) is also important, especially in RF (SHAP = 9.0).

These variables are physically linked to dust re-entrainment and filtration efficiency. Zhu et al. (2022) and Marengo et al. (2005) have shown that pressure drops strongly correlate with particulate emissions. The fact that tree-based models (XGB, RF) assign higher absolute SHAP values to these features than the ANN may reflect their ability to model non-linear interactions and threshold effects — a behavior highlighted by Lundberg et al. (2020) for tree explainability.

Temperature and Excess Air

- t smoke tower outlet (smoke tower outlet temperature) and excess air show lower but still meaningful contributions in all models.
- In the ANN, all features are assigned nearly identical SHAP values (around 1.0–1.5), suggesting that the neural network distributes importance more evenly — possibly because of its inherent non-linear smoothing or due to weight regularization.

This contrasts with Breiman (2001)'s random forests, which often produce sparser importance rankings (e.g., RF shows a gradual decline from 15.0 to 5.0 across features). The XGBoost results, following Chen & Guestrin (2016), offer an intermediate pattern (13.0 down to 9.0). The ANN's flatter importance profile may indicate lower sensitivity to individual features, which could be a strength for robustness but a weakness for interpretability, as noted in Lundberg & Lee (2017).

Model Comparison and Practical Implications

- XGBoost provides the most balanced SHAP values, with a clear ranking but modest differences between top and bottom features (13 → 9). This suggests it captures both dominant and secondary effects without over-focusing on a single variable.
- Random Forest shows the widest spread (15 → 5), implying it strongly prioritizes oven flour flow rate while downplaying excess air. This could be due to the bootstrap aggregation and random feature selection emphasized by Breiman (2001) — some features may be consistently chosen as primary splitters.

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

- Artificial Neural Network yields a compressed SHAP range (approx. 1.0 to 1.5), which may reflect the integrated gradient or linear approximation methods used in its SHAP computation. However, as Chicco et al. (2021) argue regarding R^2, absolute magnitudes matter less than relative ordering. Still, the ANN's near-equal SHAP values raise questions about feature discriminability — perhaps indicating that the network has learned redundant or highly correlated representations.

Consistency with Literature

- Su et al. (2025) found that ANNs can achieve high prediction accuracy for cement kiln emissions, but their interpretability lags behind tree-based models. Our ANN SHAP plot supports this: it identifies the same top features but with less contrast.
- Marengo et al. (2005) used neural networks to model pollutants and noted that physical variables like flow rates and temperatures are always influential — a result confirmed across our three models.
- The role of dp filter aff2 aligns with Ciobanu et al. (2021), who measured pressure drops as key indicators of filter performance in cement mills.

All three models agree that oven flour flow rate is the single most important predictor. Gas flow rate and dp filter aff2 are consistently secondary.

From a practical monitoring standpoint, this suggests that tree-based models (especially XGBoost) offer a good trade-off between predictive performance and interpretability, as argued by Lundberg et al. (2020). For control system design in cement plants, focusing first on stabilizing oven flour flow rate would yield the largest expected reduction in emissions, followed by optimizing gas flow and filter pressure drops.

Table 3. SHAP feature importance results across the three models.

Feature	SHAP XGB (XGBoost)	SHAP RF (Random Forest)	SHAP ANN (Neural Network)	Key Observations
oven flour flow rate	13.0(Highest)	15.0(Highest)	~1.0 (Top-ranked)	Most important feature across all models, confirming its dominant physical role (Ciobanu et al., 2021; Su et al., 2025).
gas flow rate	12.0(2nd)	11.0(2nd)	~1.5 (Shared top)	Consistently second-most important in tree models; ANN shows similar relative rank but compressed scale.
dp filter aff2	11.0(3rd)	9.0(4th)	~ 1.0 (Middle)	Important for filtration/pressure drop effects (Zhu et al., 2022). RF gives slightly lower rank than XGB.
temperature smoke tower outlet	10.0(4th)	7.0(5th)	~ 1.5 (Shared top)	Temperature influence is moderate; ANN ranks it higher relative to other features.
excess air	9.0(5th)	5.0(Lowest)	~1.0 (Lowest in ANN)	Least important in all models. RF downweights it most

Journal of Information Systems Engineering and Management
2026, 11(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Feature	SHAP XGB (XGBoost)	SHAP RF (Random Forest)	SHAP ANN (Neural Network)	Key Observations
				strongly; ANN compresses differences.
Importance Spread (Max – Min)	(9 → 13) 4.0	(5 → 15) 10.0	~0.5 (1.0 → 1.5)	RF shows highest contrast (prioritizes top feature). ANN shows almost uniform distribution. XGB is intermediate.
Interpretability	High (clear ranking, moderate spread)	High (very clear ranking, wide spread)	Lower (compressed values, harder to distinguish)	Tree-based models (XGB, RF) align with Lundberg et al. (2020) for explainability; ANN reflects Su et al. (2025) – accurate but less interpretable.
Suggested Reference for Behavior	Chen & Guestrin (2016) – balanced boosting	Breiman (2001) – bagging & random splits	Lundberg & Lee (2017) – gradient-based approximations, Chicco et al. (2021) – caution on magnitude comparison	

1.14. Implications for Environmental Control and Predictive Maintenance

The developed machine learning models demonstrated significant potential for improving environmental monitoring and predictive maintenance strategies in the Meftah cement plant. Accurate prediction of AFF2 dust concentration can provide an effective decision-support tool for anticipating abnormal particulate emissions and maintaining compliance with environmental regulations.

The obtained results showed that ensemble learning approaches, were capable of accurately modeling the nonlinear relationships between operating conditions and dust emissions. Such predictive capability enables plant operators to estimate future particulate concentration levels before actual exceedances occur. Consequently, corrective actions can be implemented proactively to reduce environmental risks and avoid regulatory violations. Similar applications of machine learning for industrial emission forecasting and environmental monitoring have recently demonstrated substantial improvements in process control efficiency and pollution prevention (Kim et al., 2024).

From an operational perspective, the proposed predictive models can support real-time environmental supervision systems by continuously analyzing process variables such as gas flow rate, excess air, smoke tower outlet temperature, oven flour flow rate, and differential pressure of the AFF2 filter. When the predicted dust concentration approaches regulatory thresholds, the system may generate early warning signals for operators. This predictive capability is particularly important in cement manufacturing industries where particulate emissions are strictly regulated due to their environmental and health impacts (European Environment Agency, 2023).

Moreover, the integration of predictive modeling with maintenance management strategies offers substantial advantages for preventive and predictive maintenance. In particular, abnormal increases in predicted dust concentration may indicate deterioration of filter performance, clogging phenomena, abnormal pressure drops, leakage in filtration systems, or inefficient combustion conditions. Therefore, maintenance interventions can be scheduled before severe equipment degradation or unexpected shutdowns occur. Recent studies have emphasized that machine learning-based predictive maintenance

reduces maintenance costs, improves equipment reliability, and enhances industrial sustainability (Zonta et al., 2020).

The superior performance of the Random Forest model suggests that tree-based ensemble algorithms are highly suitable for industrial environmental applications characterized by nonlinear process dynamics and noisy operational data. Their robustness and interpretability also facilitate practical implementation in industrial decision-support systems. Furthermore, SHAP analysis provides additional interpretability by identifying the most influential process variables affecting dust emissions, thereby helping engineers better understand the physical mechanisms responsible for particulate generation.

Overall, the proposed modeling framework contributes to the transition toward intelligent and sustainable cement manufacturing by combining artificial intelligence techniques with environmental management and predictive maintenance strategies. The implementation of such systems can help cement plants improve regulatory compliance, reduce particulate emissions, optimize operational efficiency, and minimize maintenance-related costs.

Conclusion

This study evaluated the performance of three machine learning models, namely Artificial Neural Network (ANN), Random Forest (RF), and XGBoost, for predicting dust concentration at the AFF2 stack of the Meftah cement plant in Blida, Algeria. The developed models were trained using several important operational parameters related to the cement manufacturing process, including gas flow rate, excess air, smoke tower outlet temperature, oven flour flow rate, and differential pressure of the AFF2 filter.

The comparative results revealed significant differences in predictive performance among the investigated models. The Random Forest model achieved the best overall performance with the lowest prediction errors and the highest coefficient of determination (MSE = 8.854, RMSE = 2.976, MAE = 1.941, and R^2 = 0.750). These results indicate that the RF model was the most effective in capturing the nonlinear relationships between process variables and particulate emissions. The strong predictive capability of Random Forest can be attributed to its ensemble learning mechanism, which improves robustness and reduces overfitting in complex industrial datasets (Breiman, 2001).

The XGBoost model also demonstrated satisfactory predictive performance, achieving an R^2 value of 0.736 with relatively low prediction errors (RMSE = 3.055 and MAE = 2.097). Although slightly less accurate than Random Forest, XGBoost remained highly competitive and showed strong ability to model nonlinear process behavior. This confirms the effectiveness of gradient boosting techniques in industrial regression applications (Chen & Guestrin, 2016).

In contrast, the ANN model produced the lowest predictive accuracy among the three approaches, with an R^2 value of 0.645 and the highest error values (RMSE = 3.547 and MAE = 2.496). Despite its ability to learn nonlinear relationships, the ANN model appeared more sensitive to data variability and may require larger datasets and more extensive hyperparameter optimization to achieve better generalization performance (Goodfellow et al., 2016).

The graphical analyses, including learning curves, prediction-versus-actual plots, and residual distributions, further confirmed the superiority of the ensemble tree-based models over the ANN model. In particular, the Random Forest model showed the closest agreement between predicted and measured dust concentrations, as well as the most stable residual distribution around the zero-error line.

Furthermore, SHAP explainability analysis provided valuable insights into the influence of operational variables on dust concentration prediction. The interpretability of the developed models represents an important advantage for industrial applications, allowing plant engineers to better understand the factors affecting particulate emissions and to improve process control strategies.

Overall, the obtained results demonstrate that machine learning models can serve as efficient tools for environmental monitoring and predictive maintenance in cement manufacturing industries. Among the investigated approaches, Random Forest proved to be the most suitable model for predicting dust emissions under the studied operating conditions. The implementation of such predictive systems may help cement plants anticipate regulatory exceedances, optimize filtration system performance, reduce environmental pollution, and improve operational sustainability.

References

[1] Al-Taai, S. R., Azize, N. M., Thoeny, Z. A., et al. (2023). "XGBoost Prediction Model Optimized with Bayesian for the Compressive Strength of Eco-Friendly Concrete." *Applied Sciences*, 13(15), 8889.

[2] ASJP (2025). "Artificial Intelligence as a Catalyst for Sustainability and Energy Transition: Opportunities, Challenges, and Future Prospects for Emerging Economies, Case of Algeria." *Algerian Scientific Journal Platform*.

[3] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[4] Breiman, Leo (2001). "Random Forests." *Machine Learning*, 45(1), 5–32.

[5] Chen, Tianqi, & Guestrin, Carlos (2016). "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

[6] European Environment Agency (2023). *Industrial Pollution in Europe: Cement Industry Environmental Performance Report*.

[7] Elhishi, S., Elashry, A. M., & El-Metwally, S. (2023). "Unboxing Machine Learning Models for Concrete Strength Prediction Using XAI." *Scientific Reports*, 13, 19892.

[8] Fayaz, S. J., & Sheikh, N. D. (2026). "A Multi-Plant Machine Learning Framework for Emission Prediction, Forecasting, and Control in Cement Manufacturing." *arXiv*, 2604.19903.

[9] Goodfellow, Ian, Bengio, Yoshua, & Courville, Aaron (2016). *Deep Learning*. MIT Press.

[10] Kim, J. H., Lee, D. H., Mendoza, J. A., & Lee, M. Y. (2024). "Applying Machine Learning Random Forest (RF) Method in Predicting Cement Products with Co-Processing of Input Materials: Optimizing the Hyperparameters." *Environmental Research*, 248, 118300.

[11] MDPI (2023). "Application of ANN, XGBoost, and Other ML Methods to Forecast Air Quality." *Sustainability*, 15(6), 5341.

[12] MDPI (2025). "A Comparison of Machine Learning-Based Approaches in Estimating Surface PM2.5 Concentrations." *Atmosphere*, 16(1).

[13] MDPI (2025). "Predicting the Concentration Levels of PM2.5 and O3 for Highly Urbanized Areas Based on Machine Learning Models." *Sustainability*, 17(20).

[14] PLOS One (2025). "Machine Learning-Based Forecasting of Air Quality Index: A Comparative Approach with XGBoost and LightGBM." *PLOS One*.

[15] ResearchGate (2025). "Machine Learning Algorithms for Predicting Air Quality Index: A Case Study in Urban and Industrial Zones." *Periodicals of Engineering and Natural Sciences (PEN)*, 13(2), 425–434.

[16] ResearchGate (2025). "Prediction of Dust Concentrations in a Cement Plant Using Artificial Neural Networks: A Process Data-Based Approach." *Journal of Environmental Control*.

[17] Safhi, A. E. M., Dabiri, H., Soliman, A., & Khayat, K. (2023). "Prediction of Self-Consolidating Concrete Properties Using XGBoost Machine Learning Algorithm." *Construction and Building Materials*, 408, 133560.

[18] Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research*, 15, 1929–1958.

[19] Sustainability (2025). "Predicting the Concentration Levels of PM2.5 Based on Machine Learning Models." *Sustainability*, 17(20).

[20] Zonta, T., da Costa, C. A., da Rosa Righi, R., de Lima, M. J., da Trindade, E. S., & Li, G. P. (2020). "Predictive Maintenance in the Industry 4.0: A Systematic Literature Review." *Computers & Industrial Engineering*, 150, 106889.