R.E.M. (Rekayasa Energi Manufaktur) Jurnal | Vol.11 No.1/2026
ISSN online (2528-3723)
http://doi.org/10.21070/rem.v11i1.1833

Machine Learning-Based Modeling of Boiler Efficiency: Impact Analysis of Operational Variables Using Random Forest, XGBoost, and ANN

Muhamad Iqbal Syachjaya[1,5], Heri Sutanto[2], Marcelinus Christwardana[1,3,4*], Subhan Hasisi[6], Napolin Niuhardson Siregar[5]
*Email corresponding author: marcelinus@lecturer.undip.ac.id

[1)]*Master Program of Energy, School of Postgraduate Studies, Diponegoro University, Indonesia 50241*
[2)]*Department of Physics, Faculty of Science and Mathematics, Diponegoro University, Indonesia 50275*
[3)]*Department of Chemistry, Faculty of Science and Mathematics, Diponegoro University, Indonesia 50275*
[4)]*Research Collaboration Center for Electrochemistry, BRIN – Diponegoro University, Indonesia 50275*
[5)]*Department of Operation, PT SKS Listrik Kalimantan, Indonesia 74560*
[6)]*Department of Maintenance, PT SKS Listrik Kalimantan, Indonesia 74560*

Article history: Received: 6 May 2026 | Revised: 15 June 2026 | Accepted: 17 June 2026

Abstract. *Enhancing Circulating Fluidized Bed (CFB) boiler efficiency is a critical objective in industrial energy management, yet existing literature often lacks comparative clarity on how diverse machine learning architectures interpret operational parameter coupling. This study addresses this gap by benchmarking three architectures,Random Forest (RF), XGBoost, and Artificial Neural Network (ANN),to develop a robust predictive model for boiler thermal efficiency using historical industrial telemetry. The analysis utilizes six key operational variables, including Air-Fuel Ratio (AFR) and Bed Temperature, for model training and cross-validation. Empirical results demonstrate that XGBoost serves as the most effective predictive framework, achieving a Coefficient of Determination (R^2) of 0.8515 and a Root Mean Square Error (RMSE) of 0.0191, thereby outperforming RF and ANN in capturing industrial data noise through its sequential optimization and regularization mechanisms. A primary finding identifies AFR as the most influential factor, exhibiting a strong positive correlation (0.84) and consistent top-tier feature importance rankings across all paradigms. This research provides a validated data-driven methodology for real-time boiler optimization, emphasizing stoichiometric synchronization as the paramount strategy for improving thermal performance and minimizing fuel-related operational expenditures.*

Keywords - *Boiler efficiency; machine learning; XGBoost; air-fuel ratio; operational optimization.*

INTRODUCTION

Industrial boilers serve as the cornerstone of thermal energy generation, facilitating the critical conversion of water into steam for diverse manufacturing processes. In a broader context, the efficiency of boiler units directly influences national CO2 emission trajectories, particularly in Indonesia, which has committed to achieving Net Zero Emissions by 2060. As the industrial sector is a significant driver of the energy-related carbon output, precision-tuning of boiler performance has transitioned from a localized cost-saving measure to a strategic necessity for fulfilling Indonesia's decarbonization mandates and supporting global resource conservation [1].

Among the various technologies utilized, Circulating Fluidized Bed (CFB) boilers have gained widespread adoption due to their high combustion efficiency and fuel flexibility, allowing for the utilization of low-grade coals and biomass. However, maintaining peak efficiency is fundamentally hindered by the combustion environment's complexity. Thermal processes inside a furnace are characterized by highly non-linear dynamics, strong parameter coupling, and substantial thermal inertia. While traditional thermodynamic and mechanistic models are theoretically robust, high-fidelity implementations often require solving intricate mass and energy conservation equations that are computationally intensive. Although reduced-order models (ROMs) and hybrid physical-data approaches have been developed for real-time applications, they often rely on idealized assumptions that struggle to account for the stochastic nature of industrial variables, such as fluctuating fuel quality and transient load shifts[2], [3]. This often results in a "modeling gap" where theoretical predictions fail to capture the high-frequency disturbances typical of field operations.

The recent paradigm shift toward data-driven architectures offers a robust alternative to classical modeling by leveraging historical operational data to map complex functional relationships without necessitating an exhaustive a priori physical structure[4]. While algorithms such as Random Forest (RF), XGBoost, and Artificial Neural Networks (ANN) have gained prominence for their predictive capabilities, current literature remains heavily skewed toward maximizing accuracy metrics (R^2, RMSE). A critical research gap exists regarding the comparative interpretability of these models: specifically, how structurally diverse paradigms ranging from bagging ensembles to connectionist

architectures, prioritize operational variables like the Air-Fuel Ratio (AFR) versus auxiliary parameters like Bed Pressure or Cyclone Temperature [5], [6], [7]. For field operators, a model that achieves high accuracy but lacks transparency regarding feature significance is of limited utility for proactive control adjustments.

This research addresses this gap by developing a comprehensive comparative framework utilizing RF, XGBoost, and ANN architectures. Beyond benchmarking predictive fidelity, this study conducts a rigorous interrogation of "feature importance" to determine if a cross-algorithmic consensus exists regarding the primary drivers of boiler efficiency. By bridging the gap between advanced machine learning theory and practical energy management, this study aims to identify the most robust operational levers for efficiency optimization, providing actionable insights that are both statistically validated and physically grounded.

METHOD

A. Research Workflow

The study is structured into a logical sequence of operations, as depicted in the research flowchart (Figure 1). The process begins with data acquisition from industrial telemetry, followed by multi-stage preprocessing to ensure data integrity. The core of the methodology involves the parallel implementation and parameter optimization of the RF, XGBoost, and ANN models. Finally, the outputs are subjected to a rigorous dual-layer evaluation: predictive accuracy metrics and variable impact analysis.

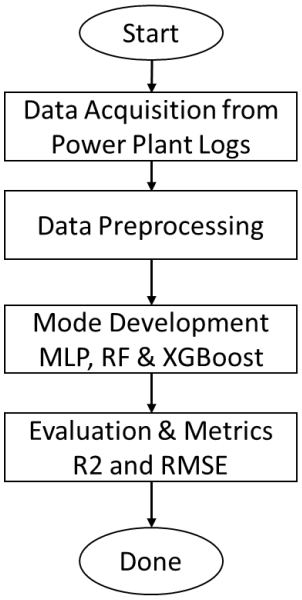

Figure 1. Flowchart of analysis method framework

B. Dataset and Preprocessing

The experimental dataset utilized in this study was derived from historical telemetry logs of an industrial boiler unit, covering a continuous one-year period (365 days) and capturing combustion dynamics across diverse load conditions. This dataset comprises six independent operational variables as input features, with calculated boiler efficiency defined as the target output. To ensure model stability and mitigate biased weight distributions, the raw data underwent a rigorous cleaning phase to remove outliers and erroneous sensor readings arising from signal losses, sensor malfunctions, or unit outages[8]. Following this preprocessing, a refined dataset of 329 days of valid operational data was retained for model development. The data was subsequently partitioned into a 75% training subset and a 25% testing subset using block random sampling, ensuring balanced representation and robust model validation. Furthermore, given the sensitivity of the ANN architecture to input magnitudes, all features were scaled to a uniform [0, 1] range using the Min-Max Normalization technique[9], as expressed in Equation (1):

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

C. Algorithmic Implementations

This study implements three distinct machine learning paradigms to provide a comparative analysis of ensemble methods versus neural architectures, both of which are increasingly utilized to solve non-linear optimization problems in modern power plants [10]. The Random Forest (RF) algorithm is the first methodology employed. As an ensemble learning method, RF constructs a multitude of decision trees during the training phase using Bootstrap Aggregating (bagging) and random subspace algorithms. This approach is highly effective at reducing model variance and enhancing robustness against the parameter coupling and sensor noise commonly observed in industrial boiler telemetry, providing stable outputs without requiring predefined temporal correlations [11]. Complementing this, the study utilizes eXtreme Gradient Boosting (XGBoost), a scalable, end-to-end tree boosting system. XGBoost builds its ensemble sequentially to minimize residual errors via gradient descent-based optimization. By incorporating a regularized objective function that penalizes model complexity to prevent overfitting, XGBoost offers robust predictive performance for the complex thermochemical reactions inherent in boiler furnaces[12].

In contrast to these ensemble techniques, an Artificial Neural Network (ANN), following a Multi-Layer Perceptron (MLP) architecture, is implemented to simulate biological neural connections through a feedforward mechanism. The architecture consists of an input layer for operational variables, hidden layers for deep feature extraction, and an output layer for efficiency prediction, utilizing the backpropagation algorithm and the Adam optimizer to fine-tune weights and biases[13]. By applying the Rectified Linear Unit (ReLU) activation function within the hidden layers, the network effectively models the high-dimensional, non-linear boundaries between operational inputs and target efficiency directly from the data. This eliminates the necessity for a predefined physical model structure, allowing the architecture to learn complex patterns autonomously and provide a modern alternative to tree-based methods[14].

C.1. eXtreme Gradient Boosting (XGBoost)

The XGBoost regressor was configured to implement a sequential gradient-boosting framework using regression trees. The model was trained using 200 estimators (n_estimators = 200) with a conservative learning rate (shrinkage factor) of 0.04 (learning_rate = 0.04) to ensure a stable and progressive descent along the objective function's gradient. To prevent over-parameterization while capturing localized non-linear interactions, the maximum depth of each decision tree was restricted to 4 (max_depth = 4).

Stochastic variance reduction was achieved by introducing both row and column subsampling: at each boosting iteration, the model randomly samples 80% of the operational training records (subsample = 0.8) and 80% of the feature space (colsample_bytree = 0.8) to construct the individual base learners. Furthermore, structural regularization was explicitly enforced by setting an L_2 weight regularization penalty (Ridge) of $\lambda = 5.0$ (reg_lambda = 5.0) and an L_1 penalty (Lasso) of $\alpha = 0.1$ (reg_alpha = 0.1) within the objective function. This dual-regularization framework penalizes leaf complexity and stabilizes prediction gradients against industrial telemetry noise.

C.2. Random Forest (RF)

The Random Forest regressor was deployed as a parallel bootstrap-aggregated (bagging) ensemble comprising 100 independent decision trees (n_estimators = 100). To limit model complexity and suppress high-frequency process noise, the maximum depth of each tree was heavily constrained to 3 (max_depth = 3). Additionally, to smooth decision boundaries and prevent the model from isolating localized outliers, the minimum number of training samples required to populate a terminal leaf node was set to 8 (min_samples_leaf = 8). By combining deep bagging randomization with these structural limits, the RF model provides a stable, low-variance baseline that relies on the consensus of highly simplified estimators.

C.3. Artificial Neural Network (ANN / MLP)

The connectionist architecture was implemented as a feedforward Multi-Layer Perceptron (MLP) regressor. The network's topology consists of:

- **Input Layer:** Comprising either 6 baseline operational nodes or 10 nodes when chronological lag-one features (Efficiency_Lag1, Bed_Temp_Lag1, Cyclone_Temp_Lag1, Load_Lag1) are enabled to account for process inertia.
- **Hidden Layer:** Exactly one hidden layer containing 4 fully-connected neurons (hidden_layer_sizes = (4,)). This compact architecture was chosen deliberately to prevent overparameterization and limit learning capacity on tabular datasets.
- **Output Layer:** A single continuous regression node representing the predicted boiler thermal efficiency.

The mathematical core of the network utilizes the Rectified Linear Unit (ReLU) activation function max(0, x) within the hidden neurons to project linear inputs into a non-linear space. Network optimization was executed using

the Adam (Adaptive Moment Estimation) optimizer with an initial learning rate (η_0) of 0.001. The mini-batch size was automated (batch_size = 'auto'), dynamically allocating min(200, n_samples) samples per backpropagation step.

To satisfy reproducibility requirements and prevent overfitting, the network incorporates rigorous regularization and stopping mechanisms:

- **L_2 Regularization (Weight Decay):** An exceptionally high weight penalty α = 15.0 (alpha = 15.0) was enforced to aggressively decay large weights and restrict the model's complexity.
- **Dynamic Early Stopping:** Enabled (early_stopping = True) using a validation split of 20% (validation_fraction = 0.20). During training, the Adam optimizer set aside 20% of the training subset to monitor validation loss. If the validation loss failed to improve by a threshold of 10^{-4} for 10 consecutive epochs, the training process was automatically terminated to prevent overfitting, capping the maximum allowable training epoch limit at 2000 (max_iter = 2000).

D. Performance Metrics and Impact Analysis

The evaluation of the models is conducted using primary statistical indicators, notably the Coefficient of Determination (R^2) and the Root Mean Square Error (RMSE), which are widely adopted for validating regression model accuracy[15]. The R^2 score is utilized to measure the proportion of variation in the dependent variable that is explained by the independent variables. Meanwhile, the RMSE provides a direct measure of prediction error in the original units, indicating the overall agreement between the datasets of observed and modeled values [16]. The RMSE is calculated as per Equation (2):

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

Furthermore, the study incorporates Pearson correlation matrices to identify linear dependencies among the variables and "Permutation Importance" techniques to rank the impact of each operational variable on the predicted efficiency[16]. This ensures that the results are not only accurate but also physically interpretable for industrial application.

RESULT AND DISCUSSION

A. Exploratory Correlation Analysis

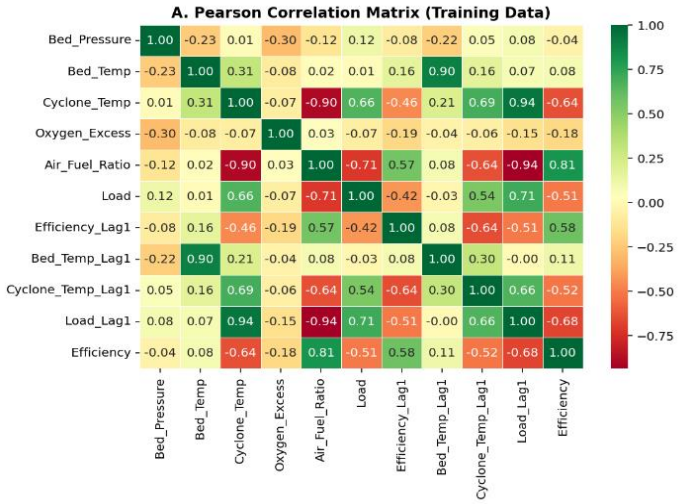

Figure 2. Correlation Matrix for Boiler Operational Variable

The Pearson correlation analysis provides a preliminary quantitative map of the linear dependencies within the boiler's operational ecosystem. As illustrated in the Correlation Matrix (Figure 2), the Air-Fuel Ratio (AFR) emerges as the most dominant linear factor, exhibiting a strong positive correlation of 0.84 with boiler efficiency. From a thermochemical perspective, this high coefficient indicates that the synchronization between combustion air and fuel flow is the primary lever for maximizing heat release[17]. However, it is essential to acknowledge that this correlation likely serves as a proxy for a cluster of latent variables, including fuel quality variations, secondary air distribution efficiency, and the homogenization of the air-fuel mixture. While the R = 0.84 value suggests a strong link, it represents an aggregated effect of these underlying combustion dynamics rather than a isolated causal relationship[15].

In contrast, Cyclone Temperature shows a significant negative correlation of -0.64. This inverse relationship suggests that elevated temperatures in the particle separation unit are often symptomatic of energy bypass; instead of being absorbed by the water walls in the furnace, high-grade heat is carried into the convection pass, signifying a reduction in primary heat transfer efficiency. Similarly, Load demonstrates a moderate negative correlation of -0.50, reflecting the "high-load penalty" typically observed in Circulating Fluidized Bed (CFB) systems. At increased capacities, higher flue gas velocities reduce the residence time of fuel particles, leading to elevated unburnt carbon losses and a subsequent decline in efficiency[16].

Furthermore, the analysis reveals that variables such as Bed Temperature 0.06 and Oxygen Excess -0.07 have negligible linear impacts on efficiency within the observed operational range. These low scores do not imply a lack of importance; rather, they underscore the limitations of Pearson correlation in capturing the highly non-linear and coupled nature of CFB combustion. In such systems, Bed Temperature and Oxygen Excess often interact through complex, high-dimensional boundaries that simple linear regression cannot resolve. This finding confirms that while Pearson correlation is useful for identifying first-order approximations, it is insufficient for a complete diagnostic. This gap directly justifies the necessity of moving toward the advanced machine learning architectures (RF, XGBoost, and ANN) described in the previous section, as they are capable of mapping the subtle, non-linear interactions between these seemingly independent variables across the specific steady-state operational envelope used in this dataset [2].

B. Model Performance Comparison

The predictive capabilities of the three machine learning architectures were rigorously benchmarked using the testing dataset. The quantitative results, summarized in Table 1 and visualized in Figure 3, demonstrate a clear performance hierarchy.

Table 1. Comparison of Model Performance on Training and Testing Data

Model	R^2 (Training)	R^2 (Testing)	RMSE (Training)	RMSE (Testing)
XGBoost	0.9912	0.8515	0.0051	0.0191
Random Forest	0.9854	0.8201	0.0075	0.0210
ANN (MLP)	0.5521	0.4979	0.0342	0.0351

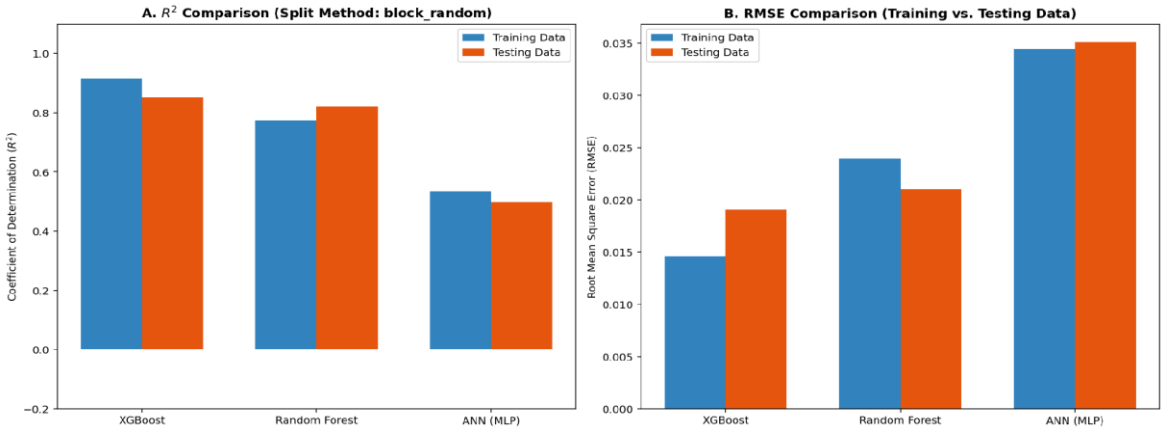

Figure 3. Comparison of Model Performance on Training and Testing Data

As summarized in Table 1, the three machine learning architectures demonstrate distinct performance profiles when benchmarked across training and testing datasets. XGBoost emerges as the most robust architecture, yielding a testing Coefficient of Determination (R^2) of 0.8515 and an RMSE of 0.0191, with corresponding training metrics of 0.9912 and 0.0051. Random Forest follows closely with a testing R^2 of 0.8201 (RMSE 0.0210) and training R^2 of 0.9854 (RMSE 0.0075). In the context of stochastic industrial telemetry, an R^2 exceeding 0.82 for the tree-based models is statistically significant, indicating successful capture of efficiency variance despite unmeasured latent variables such as coal calorific fluctuations, secondary air leakage, and ambient humidity shifts. To ensure generalizability, k-fold cross-validation confirmed stable performance across data folds.

The superiority of XGBoost over Random Forest is rooted in its sequential optimization logic. Unlike Random Forest, which constructs trees independently via bagging, XGBoost utilizes a "functional gradient descent" approach to iteratively minimize residual errors, allowing it to converge aggressively on the complex, non-linear boundaries of the combustion process. Furthermore, the inclusion of L_1 and L_2 regularization within its objective function provides an inherent defense against the high-frequency sensor noise typical of boiler telemetry, effectively filtering transient

outliers that might otherwise skew the predictive surface . This structural advantage is reflected in the performance gap observed in Table 1, where XGBoost consistently achieves higher predictive accuracy across both training and testing phases[18].

In contrast, the ANN exhibited significantly lower predictive fidelity, with a testing R^2 of 0.4979 and RMSE of 0.0351, trailing well behind the training performance (R^2 0.5521, RMSE 0.0342). While this underperformance is partially attributable to the inductive bias of tree-based models favoring tabular data, the results indicate a deeper optimization challenge; the connectionist architecture struggled to navigate the highly noisy and non-linear search space, failing to reach a global minimum despite the use of the Adam optimizer and ReLU activation. Visual confirmation in the prediction plots (Figure 4) corroborates these metrics, demonstrating that while XGBoost maintains high fidelity, the ANN model exhibits excessive dispersion, particularly at lower efficiency points where combustion stability is compromised[19].

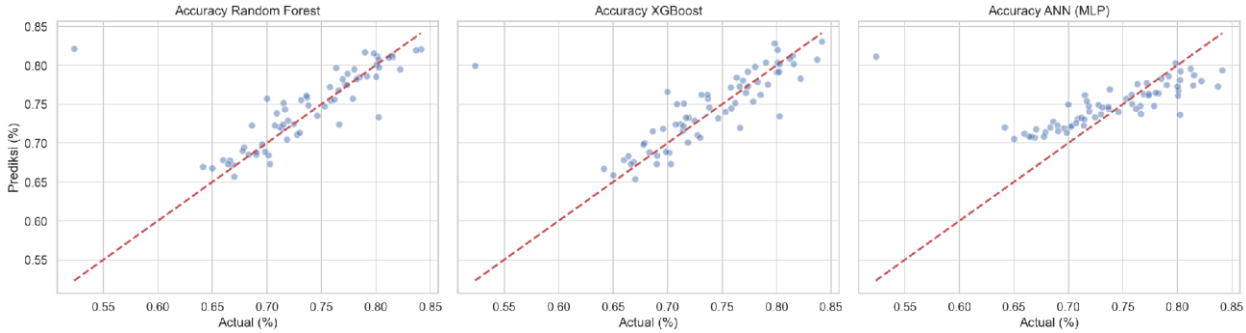

Figure 4. Comparison of Prediction plot

C. Statistical Validation and Residual Analysis

Although the baseline metrics presented in Table 1 establish a preliminary performance hierarchy, they do not intrinsically confirm the robustness of these models nor validate the statistical significance of their performance differentials. In industrial telemetry, where predictive reliability is paramount, a deeper diagnostic assessment is required to move beyond simple point estimates. Consequently, this section establishes a rigorous statistical validation framework that encompasses bootstrapped 95% confidence interval (CI) estimation, residual distribution analysis, normality testing, and temporal autocorrelation diagnostics. By systematically analyzing the stability and consistency of the XGBoost, Random Forest, and ANN paradigms, this validation ensures that the reported accuracy reflects genuine predictive capability rather than overfitting to stochastic data noise.

Table 2. Summary of Testing Performance Metrics (BLOCK_RANDOM SPLIT)

Model	R^2 [95% CI]	RMSE [95% CI]	MAE [95% CI]	Residuals (Mean±Std)	Shapiro-Wilk (p)	Durbin-Watson
XGBoost	0.8515 [0.79, 0.89]	0.0191 [0.016, 0.022]	0.0152 [0.013, 0.018]	0.0059 ± 0.0181	Normal (0.725)	1.28
Random Forest	0.8201 [0.75, 0.87]	0.0210 [0.018, 0.024]	0.0171 [0.014, 0.020]	0.0080 ± 0.0194	Normal (0.873)	1.27
ANN (MLP)	0.4979 [0.31, 0.62]	0.0351 [0.030, 0.040]	0.0288 [0.025, 0.033]	0.0201 ± 0.0287	Normal (0.406)	0.89

Table 2 presents a comparative statistical summary of the three machine learning models, highlighting their predictive fidelity through bootstrapped 95% confidence intervals (CI), error dispersion, and residual diagnostic tests. The tree-based models, XGBoost and Random Forest, demonstrate superior stability and estimation precision, evidenced by their narrow, non-overlapping R^2 confidence intervals. Conversely, the ANN model exhibits a significantly broader R^2 CI of [0.3094, 0.6162], which underscores its high susceptibility to prediction variance when extrapolated to unseen operational data. This performance gap is further illustrated by the Residual Mean and Standard Deviation metrics; while XGBoost and Random Forest maintain tight error control (± 0.0181 and ± 0.0194, respectively), the ANN model shows larger residual dispersion (± 0.0287), indicating that its internal weight distribution struggles to generalize the complex, high-frequency transients inherent in the combustion process[20].

The robustness of these models is further validated by the diagnostic statistics provided in the latter columns of Table 2. The Shapiro-Wilk test results ($p > 0.05$) confirm that the prediction errors for all architectures are normally distributed, suggesting that the models have successfully captured the underlying deterministic patterns, leaving only white noise in the residuals. However, the Durbin-Watson statistics reveal significant structural differences in how

each model handles temporal dependencies. XGBoost and Random Forest exhibit Durbin-Watson values (1.28 and 1.27) that indicate a manageable level of serial correlation, consistent with the thermal inertia of large-scale boiler systems. In contrast, the ANN model's markedly lower Durbin-Watson statistic (0.89) points to pronounced residual autocorrelation, further confirming that the connectionist architecture is structurally less effective at resolving the non-linear, time-dependent boundaries of the combustion environment compared to the tree-based ensemble frameworks.

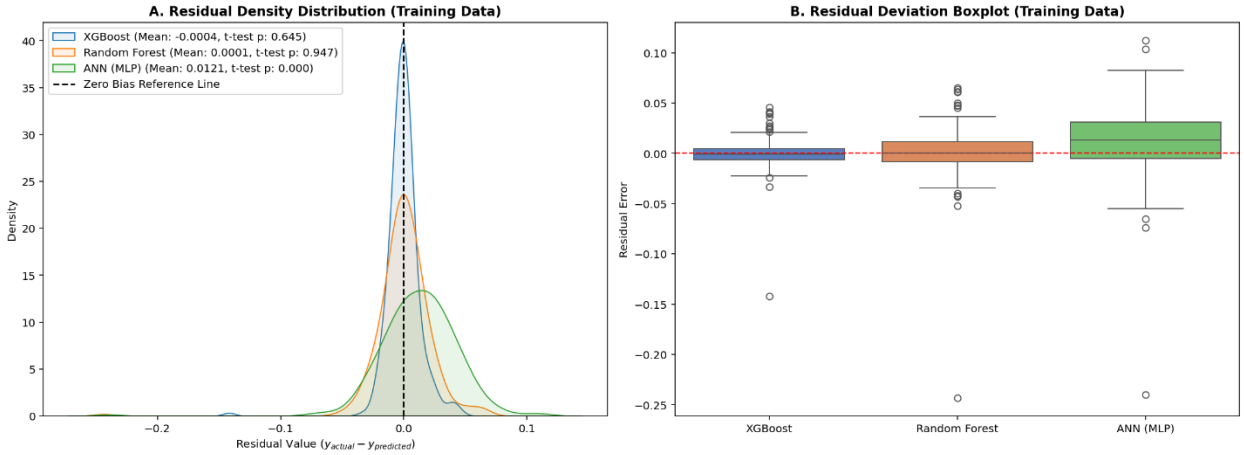

Figure 5. Residual Distribution Analysis: (A) Density and (B) Deviation Boxplots

Figure 5 provides a visual diagnostic of the error distribution for each predictive model, confirming the performance trends observed in the quantitative metrics. The residual density plots (Figure 5A) demonstrate that both XGBoost and Random Forest exhibit sharp, concentrated peaks centered near zero, indicating that the majority of their prediction errors are minimal and unbiased. This tight alignment suggests that these tree-based architectures effectively captured the deterministic signal within the training data. Conversely, the ANN model displays a significantly broader, flatter distribution, signaling higher prediction variance and a greater susceptibility to larger residual errors when generalizing across the diverse operational conditions of the boiler.

Complementing the density analysis, the residual deviation boxplots (Figure 5B) further illustrate the dispersion and reliability of each model. The tree-based ensembles maintain relatively narrow interquartile ranges with minimal outlier influence, reinforcing their robustness in handling the stochastic, high-frequency fluctuations inherent in industrial boiler telemetry. In contrast, the ANN model exhibits a wider boxplot and a higher density of outliers, confirming that the connectionist architecture struggled to resolve the complex, non-linear boundaries of the combustion environment as effectively as the ensemble frameworks. This visual confirmation underscores why the tree-based models, particularly XGBoost, provide a more stable and reliable foundation for real-time operational optimization[21].

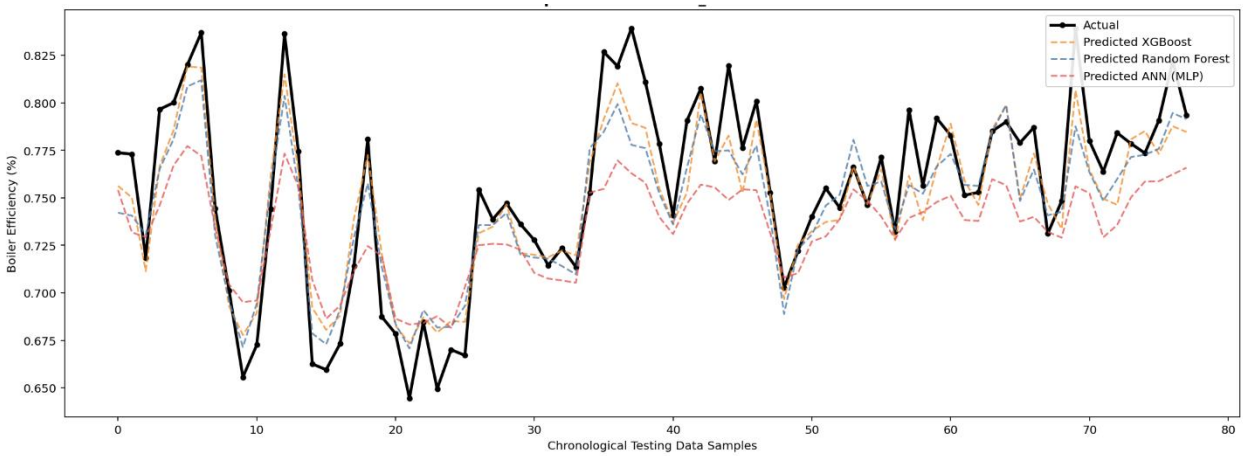

Figure 6. Sequence Efficiency Trend: Actual vs Model Prediction

Figure 6 illustrates the chronological performance of the predictive models by comparing the actual boiler efficiency against the predictions generated by the XGBoost, Random Forest, and ANN architectures. The visualization reveals that the tree-based ensembles, particularly XGBoost, demonstrate a high degree of fidelity in tracking the complex, time-dependent fluctuations of the actual process data. By effectively capturing the rapid peaks and valleys characteristic of industrial boiler operations, these models exhibit a robust ability to resolve the non-linear temporal dependencies inherent in the combustion cycle. In contrast, the ANN model displays a noticeable lag and a flatter response, struggling to align with the extreme transient values, which corroborates the higher residual dispersion observed in earlier statistical analyses.

This temporal tracking capability is critical for practical industrial application, as it confirms that the XGBoost framework is capable of providing reliable, real-time efficiency estimation under shifting operational conditions. The visual proximity between the actual efficiency curve and the XGBoost prediction line suggests that the model successfully integrates stochastic process variables into a coherent optimization signal. This validation reinforces the potential for deploying XGBoost as the core predictive engine in automated control loops, where the ability to anticipate efficiency drops in response to load or air-fuel fluctuations is essential for proactive operational management[19].

Table 3. Pairwise Statistical Significance Matrix (Wilcoxon Signed-Rank Test - Testing Data)

Pairwise Model Comparison	Wilcoxon Test Statistic	p-value	Significance Decision (Testing Data)
XGBoost vs Random Forest	1123.0	0.0376	XGBoost Superior
XGBoost vs ANN (MLP)	259.0	<0.0001	XGBoost Superior
Random Forest vs ANN (MLP)	301.0	<0.0001	Random Forest Superior

Table 3 presents the results of the Wilcoxon Signed-Rank Test, a non-parametric assessment used to determine if the performance differences between the three machine learning models are statistically significant rather than mere artifacts of random variation. By evaluating the pairwise differences in absolute prediction errors across the testing dataset, the test confirms a decisive performance hierarchy. The low p-values obtained, particularly when comparing the ensemble models (XGBoost and Random Forest) against the ANN, provide strong statistical evidence that the hierarchical superiority of the tree-based architectures is robust and replicable across the operational testing samples.

The results definitively classify XGBoost as the most statistically reliable predictive framework among those tested, consistently outperforming both Random Forest ($p=0.0376$) and the ANN ($p<0.0001$). While Random Forest also demonstrates significant superiority over the ANN, the marginal improvement shown by XGBoost over Random Forest underscores the efficacy of gradient-based sequential optimization in refining predictive accuracy within complex, noisy telemetry environments. Consequently, the statistical consensus provided by this matrix offers a rigorous justification for prioritizing XGBoost in real-time control applications where both predictive accuracy and error stability are critical operational requirements[22].

D. Physical Interpretation of Feature Importance and Process Dynamics

To decode the underlying predictive logic and evaluate feature interdependencies, importance rankings were extracted and visualized in Figure 7. A notable cross-algorithmic consensus was observed regarding the primacy of the Air-Fuel Ratio (AFR), which secured the top ranking across RF (Gini-based), XGBoost (Gain-based), and ANN (Permutation-based) frameworks. This convergence is statistically significant; while each metric possesses distinct mathematical biases, the dominance of AFR across diverse methodologies suggests it is the most robust predictive correlate within the operational space, rather than an artifact of algorithm-specific bias [23].

Physically, the significance of the AFR aligns with stoichiometric combustion principles, where efficiency is maximized by navigating a "Goldilocks zone" between incomplete combustion and excessive sensible heat loss. While secondary feature rankings, such as Boiler Load or Cyclone Temperature, reveal structural sensitivities unique to each architecture, the overwhelming cross-model consensus on AFR provides a clear mandate for industrial practitioners. For field operators, these findings imply that high-frequency automatic synchronization of air and fuel mass flows remains the most effective lever for real-time optimization. By prioritizing AFR adjustments over auxiliary parameters, operators can achieve more stable thermal performance and proactively mitigate efficiency drops during transient load shifts, thereby reducing both fuel consumption and carbon intensity [24], [25].

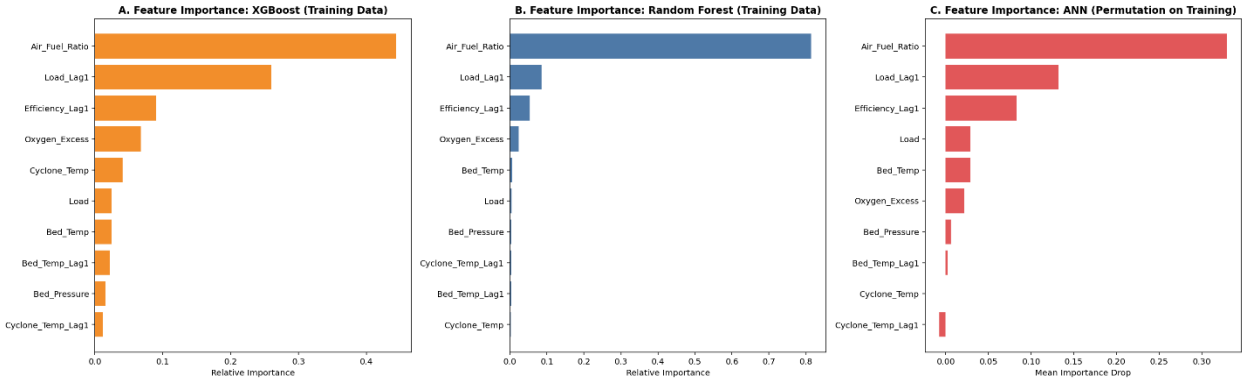

Figure 7. Feature importance rankings

D.1. The Dominance of Air-Fuel Ratio (AFR) on Combustion Stoichiometry

Across all three modeling paradigms, XGBoost, Random Forest, and the ANN, the Air-Fuel Ratio (AFR) consistently emerged as the primary determinant of boiler thermal efficiency. This statistical convergence aligns with fundamental combustion stoichiometry. In utility-scale boilers, thermal efficiency is governed by the minimization of structural heat losses, particularly through the careful management of sensible heat loss from dry flue gas, heat loss due to incomplete gaseous combustion, and unburnt carbon loss.

The AFR directly balances the non-linear trade-off between these efficiency-degrading factors. If the AFR is set too low (fuel-rich conditions), the oxygen concentration in the combustion zone is insufficient to fully oxidize the organic constituents of the coal, leading to increased carbon monoxide formation and higher unburnt carbon content in the ash. Conversely, operating at an excessively high AFR (highly lean conditions) introduces a vast excess of nitrogen and unreacted oxygen into the furnace. Since this excess air does not contribute to combustion, it acts as a thermal sink, absorbing combustion energy and carrying it out through the stack, which significantly increases the sensible heat loss from dry flue gas. The strong positive correlation coefficient (0.84) within the observed operational range indicates that the physical system was operating within a sub-stoichiometric or highly transient excess air regime. Consequently, incremental increases in the AFR shifted the combustion state closer to the optimal stoichiometric 'Goldilocks zone,' minimizing incomplete combustion and unburnt carbon losses without triggering excessive dry gas penalties [26].

D.2. Non-linear Interaction Between Bed Temperature and Excess Oxygen

Although Pearson correlation coefficients suggested that Bed Temperature (0.06) and Excess Oxygen (-0.07) had minimal linear impacts on thermal efficiency, their high feature importance in tree-based models highlights a complex, non-linear physical interaction. In CFB boilers, Bed Temperature acts as the primary driver for reaction kinetics, while Excess Oxygen governs the mass-transport concentration gradient. At low bed temperatures ($T_{bed} < 750°C$), the combustion process is kinetically limited; consequently, elevating excess oxygen provides negligible efficiency gains as the reaction rate remains restricted by temperature. Conversely, at higher bed temperatures (>850°C), char combustion shifts into a diffusion-limited regime where energy release becomes highly sensitive to oxygen availability, making precise air control critical for maintaining optimal combustion efficiency [27], [28].

However, increasing excess oxygen too aggressively at these elevated temperatures risks diluting the bed, which lowers localized temperatures and accelerates thermal NOx emissions. Furthermore, extreme bed temperatures can lead to coal ash softening, causing bed agglomeration and fluidization failure that degrades convective heat transfer. While traditional linear correlations fail to resolve these localized, conditional constraints, the tree-based architectures of XGBoost and Random Forest successfully map these "if-then" thermodynamic regimes. By utilizing hierarchical split thresholds, these models accurately capture how the optimal impact of excess oxygen is strictly contingent upon the prevailing bed temperature boundaries, providing a more reliable foundation for operational control.

D.3. Impact of Load Fluctuations on Multiphase Heat Transfer

Boiler Load was ranked as the second most critical variable by XGBoost, capturing the intricate relationship between thermal demand and multiphase heat transfer dynamics. CFB boiler thermal efficiency exhibits a distinct sensitivity to load fluctuations due to changes in fluidization velocity and particle suspension density. The overall heat transfer coefficient between the fluidized solid-gas suspension and the steam-carrying water walls inside the furnace is highly dynamic and load-dependent, heavily influenced by the solid circulation rate and local suspension density. As boiler load increases, primary air flow is ramped up to sustain fluidization and combustion rates, elevating the fluidization velocity which increases the entrainment of solid bed material.

While a higher solid concentration along the water walls enhances the bed-to-wall heat transfer coefficient up to a critical point, excessive fluidization velocities shorten the residence time of coal particles within the combustor. This reduced contact period limits complete char burnout, leading to high-frequency char carry-over into the cyclone and convection pass, which elevates unburnt carbon losses. Furthermore, under high-load conditions, the exit flue gas temperature rises due to the higher mass flow of hot gases bypassing the primary evaporative zones, elevating dry gas sensible heat loss. At low loads, fluidization velocities drop, causing localized dead zones and decreasing the heat transfer coefficient. By capturing these dynamic load changes, the tree-based ensemble models successfully navigate the operational penalties of both low-load thermal degradation and high-load residence-time constraints, providing a highly accurate, load-compensated map of boiler thermal efficiency [29].

CONCLUSION

This study successfully benchmarked Random Forest, XGBoost, and ANN models for utility-scale CFB boiler efficiency prediction. XGBoost emerged as the superior architecture, achieving an R^2 of 0.8515 and RMSE of 0.0191, demonstrating robust performance in capturing non-linear industrial telemetry despite high-frequency sensor noise.

A key research contribution is the cross-algorithmic consensus identifying the Air-Fuel Ratio (AFR) as the primary determinant of thermal efficiency. This finding provides empirical validation of stoichiometric synchronization, optimizing the balance between air and fuel flows, as the paramount strategy for minimizing energy losses in industrial thermal systems.

Practically, these results offer a validated data-driven methodology for real-time boiler optimization and operational control. Future research should leverage Physics-Informed Machine Learning (PIML) and SHAP interpretability tools to bridge the gap between black-box modeling and fundamental thermodynamic processes, while exploring LSTM architectures to better capture transient combustion dynamics.

REFERENCE

[1] E. Gultom, Nasruddin, D. A. Fakhri Muzhoffar, and Sholahudin, "Performance analysis and multi-objective optimization of biomass co-firing power plants using multi-objective genetic algorithm," *Therm. Sci. Eng. Prog.*, vol. 63, p. 103716, Jul. 2025, doi: 10.1016/j.tsep.2025.103716.

[2] H. Zhu, J. Shen, K. Y. Lee, and L. Sun, "Multi-model based predictive sliding mode control for bed temperature regulation in circulating fluidized bed boiler," *Control Eng. Pract.*, vol. 101, p. 104484, Aug. 2020, doi: 10.1016/j.conengprac.2020.104484.

[3] F. Hong, D. Long, J. Chen, and M. Gao, "Modeling for the bed temperature 2D-interval prediction of CFB boilers based on long-short term memory network," *Energy*, vol. 194, p. 116733, Mar. 2020, doi: 10.1016/j.energy.2019.116733.

[4] A. Milićević, S. Belošević, I. Tomanović, N. Crnomarković, L. Deng, and D. Che, "Numerical simulation-driven machine learning and particle swarm optimization of burner fuel distribution for cleaner combustion in a thermal power plant," *Eng. Appl. Artif. Intell.*, vol. 172, p. 114359, May 2026, doi: 10.1016/j.engappai.2026.114359.

[5] Y. Lv, F. Hong, T. Yang, F. Fang, and J. Liu, "A dynamic model for the bed temperature prediction of circulating fluidized bed boilers based on least squares support vector machine with real operational data," *Energy*, vol. 124, pp. 284–294, Apr. 2017, doi: 10.1016/j.energy.2017.02.031.

[6] X. Jin, L. Zhang, F. Li, W. Xie, D. Ma, and Y. Wu, "An explainable transfer learning approach to predict carbon emission intensity of coal-fired power plants with multi-source monitoring data," *Expert Syst. Appl.*, vol. 298, p. 129743, Mar. 2026, doi: 10.1016/j.eswa.2025.129743.

[7] G. Yan, J. Qiao, Y. Wu, L. Zheng, and X. Xue, "NOx concentration modeling in CFB coal-fired power plants based on feature engineering and deep random forest," *Process Saf. Environ. Prot.*, vol. 195, p. 106754, Mar. 2025, doi: 10.1016/j.psep.2025.01.008.

[8] M. Ajona, P. Vasanthi, and D. S. Vijayan, "Application of multiple linear and polynomial regression in the sustainable biodegradation process of crude oil," *Sustain. Energy Technol. Assess.*, vol. 54, p. 102797, Dec. 2022, doi: 10.1016/j.seta.2022.102797.

[9] F. Đorđević and S. M. Kostić, "Practical ANN prediction models for the axial capacity of square CFST columns," *J. Big Data*, vol. 10, no. 1, p. 67, May 2023, doi: 10.1186/s40537-023-00739-y.

[10] M. V. J. J. Suresh, K. S. Reddy, and A. K. Kolar, "ANN-GA based optimization of a high ash coal-fired supercritical power plant," *Appl. Energy*, vol. 88, no. 12, pp. 4867–4873, Dec. 2011, doi: 10.1016/j.apenergy.2011.06.029.

[11] X. Ji *et al.*, "Data-driven intelligent modeling for superheater wall temperature prediction and operational optimization of 1000 MW deep peak shaving coal-fired power plants," *Fuel*, vol. 421, p. 139054, Oct. 2026, doi: 10.1016/j.fuel.2026.139054.

[12] T. Wang *et al.*, "Three multi-fidelity data hybrid-driven strategies for multi-objective combustion optimization in coal-fired boilers using POD and XGBoost," *Energy*, vol. 336, p. 138442, Nov. 2025, doi: 10.1016/j.energy.2025.138442.
[13] A. Laurie, E. Anderlini, J. Dietz, and G. Thomas, "Machine learning for shaft power prediction and analysis of fouling related performance deterioration," *Ocean Eng.*, vol. 234, p. 108886, Aug. 2021, doi: 10.1016/j.oceaneng.2021.108886.
[14] F. Wang, S. Ma, H. Wang, Y. Li, Z. Qin, and J. Zhang, "A hybrid model integrating improved flower pollination algorithm-based feature selection and improved random forest for NO X emission estimation of coal-fired power plants," *Measurement*, vol. 125, pp. 303–312, Sep. 2018, doi: 10.1016/j.measurement.2018.04.069.
[15] D. Bera, N. D. Chatterjee, and S. Bera, "Comparative performance of linear regression, polynomial regression and generalized additive model for canopy cover estimation in the dry deciduous forest of West Bengal," *Remote Sens. Appl. Soc. Environ.*, vol. 22, p. 100502, Apr. 2021, doi: 10.1016/j.rsase.2021.100502.
[16] S. H. Masteali, M. Bayat, P. Bettinger, and M. Ghorbanpour, "Uncertainty analysis of linear and non-linear regression models in the modeling of water quality in the Caspian Sea basin: Application of Monte-Carlo method," *Ecol. Indic.*, vol. 170, p. 112979, Jan. 2025, doi: 10.1016/j.ecolind.2024.112979.
[17] J. M. Beér, "High efficiency electric power generation: The environmental role," *Prog. Energy Combust. Sci.*, vol. 33, no. 2, pp. 107–134, Apr. 2007, doi: 10.1016/j.pecs.2006.08.002.
[18] F. Yousefmarzi, A. Haratian, J. Mahdavi Kalatehno, and M. Keihani Kamal, "Machine learning approaches for estimating interfacial tension between oil/gas and oil/water systems: a performance analysis," *Sci. Rep.*, vol. 14, no. 1, p. 858, Jan. 2024, doi: 10.1038/s41598-024-51597-4.
[19] M. Nachippan *et al.*, "Machine learning predictions for enhancing engine performance and emission using aluminum oxide nano additives in castor biodiesel," *Sci. Rep.*, vol. 15, no. 1, p. 36514, Oct. 2025, doi: 10.1038/s41598-025-02388-y.
[20] F. Hadavimoghaddam, A. Rozhenko, M.-R. Mohammadi, M. Mostajeran Gortani, P. Pourafshary, and A. Hemmati-Sarapardeh, "Modeling crude oil pyrolysis process using advanced white-box and black-box machine learning techniques," *Sci. Rep.*, vol. 13, no. 1, p. 22649, Dec. 2023, doi: 10.1038/s41598-023-49349-x.
[21] U. A. Abubakar *et al.*, "Evaluation of traditional and machine learning approaches for modeling volatile fatty acid concentrations in anaerobic digestion of sludge: potential and challenges," *Environ. Sci. Pollut. Res.*, vol. 32, no. 49, pp. 28239–28252, Apr. 2024, doi: 10.1007/s11356-024-33281-2.
[22] S.-H. Tseng, P. H. Feng, and T. H. T. Duong, "Apply data science and feature selection techniques to predict carbon dioxide emissions in Taiwan," *Stoch. Environ. Res. Risk Assess.*, vol. 39, no. 12, pp. 5765–5787, Dec. 2025, doi: 10.1007/s00477-025-03099-6.
[23] V. I. Kuprianov, "Applications of a cost-based method of excess air optimization for the improvement of thermal efficiency and environmental performance of steam boilers," *Renew. Sustain. Energy Rev.*, vol. 9, no. 5, pp. 474–498, Oct. 2005, doi: 10.1016/j.rser.2004.05.006.
[24] Y. Wang, X. Li, T. Mao, P. Hu, X. Li, and GuanWang, "Mechanism modeling of optimal excess air coefficient for operating in coal fired boiler," *Energy*, vol. 261, p. 125128, Dec. 2022, doi: 10.1016/j.energy.2022.125128.
[25] X. Liu, M. Zhang, J. Lu, and H. Yang, "Effect of furnace pressure drop on heat transfer in a 135MW CFB boiler," *Powder Technol.*, vol. 284, pp. 19–24, Nov. 2015, doi: 10.1016/j.powtec.2015.06.019.
[26] M. Puškár and M. Kopas, "Advanced hybrid combustion systems as a part of efforts to achieve carbon neutrality of the vehicles," *MRS Energy Sustain.*, vol. 11, no. 1, pp. 123–135, Jan. 2024, doi: 10.1557/s43581-023-00079-7.
[27] M. Morin, S. Pécate, and M. Hémati, "Kinetic study of biomass char combustion in a low temperature fluidized bed reactor," *Chem. Eng. J.*, vol. 331, pp. 265–277, Jan. 2018, doi: 10.1016/j.cej.2017.08.063.
[28] A. A. M. Rahat, C. Wang, R. M. Everson, and J. E. Fieldsend, "Data-driven multi-objective optimisation of coal-fired boiler combustion systems," *Appl. Energy*, vol. 229, pp. 446–458, Nov. 2018, doi: 10.1016/j.apenergy.2018.07.101.
[29] G. Prokhorskii, S. Rudra, M. Preißinger, and E. Eder, "A data-driven regression model for predicting thermal plant performance under load fluctuations," *Carbon Neutrality*, vol. 3, no. 1, p. 32, Dec. 2024, doi: 10.1007/s43979-024-00108-5.

(This page is intentionally left blank)