

## PREDICTION AND ANALYSIS OF INDUSTRIAL ENERGY CONSUMPTION USING MACHINE LEARNING TECHNIQUES

Dr. Mosebetsi Johannes Leotlela<sup>1\*</sup>, Bhavarlal Kesharimalji Chordiya<sup>2</sup>

<sup>1\*</sup>Researcher University of the Witwatersrand, School of Physics

<sup>2</sup>Professor Dr. D.Y. Patil Institute of Master of Computer Applications and Management

### **Abstract**

*Energy use is a critical issue for efficiency and sustainable production in particular industries where energy use is high. Predicting and understanding energy demand can aid energy monitoring, planning and efficiency. This study aimed to analyze and predict industrial energy consumption using machine learning techniques while identifying the key variables influencing energy demand. A quantitative secondary data analysis was conducted using 35,040 observations of electrical, temporal, and operational variables from an industrial steel production context. CO<sub>2</sub> emissions were excluded to avoid redundancy. Descriptive statistics, correlation analysis, and temporal trend evaluation were performed. Three regression models (Linear Regression, Decision Tree and Random Forest) were trained on an 80:20 split. We used MAE, RMSE and R<sup>2</sup> to evaluate the model and the Random Forest feature importance to determine important features. Energy consumption exhibited a right-skewed distribution with clear hourly operational patterns. The Random Forest model achieved the best performance, with MAE = 0.2917, RMSE = 0.8864, and R<sup>2</sup> = 0.9993. Lagging reactive power emerged as the dominant predictor, accounting for 87.03% of feature importance, followed by lagging power factor. These findings demonstrate that machine learning models can effectively predict industrial energy consumption while providing interpretable insights into key electrical drivers. The results highlight the importance of reactive power and power factor management for improving industrial energy efficiency.*

**Keywords:** industrial energy consumption, machine learning, Random Forest, reactive power, energy efficiency

---

DOI: 10.69980/as.v12i2.6673

Received:30-12-2025 | Revised:02-03-2026 | Accepted:09-04-2026 | Published:27-04-2026

Copyright: © 2026 The Author(s). This work is licensed under a [CC BY-NC-ND 4.0](https://creativecommons.org/licenses/by-nc-nd/4.0/) International License.

## 1. Introduction

Industrial energy consumption is an important aspect of sustainability, efficiency, and environmental considerations. The steel industry, in particular, is known to be one of the highest energy consumers, contributing substantially to global industrial energy consumption and environmental impacts. Reducing energy use and emissions has become a critical focus for enhancing the sustainability of steel manufacturing operations, driven by tightened regulations and global climate change targets (Holappa, 2020). The intricacies of steel production processes, including their multiple stages and variable operating conditions, add to the challenges of energy management. This is further complicated by the interactions between energy use, productivity and environmental performance across the steel value chain (Conejo et al., 2020). Energy efficiency in these settings is affected by technical and organizational elements. These factors include the efficiency of equipment, operational practices and the deployment of monitoring systems. But factors such as a lack of real-time data integration, weak analytical skills, and poor integration of energy management practices can limit their performance (Ahmad et al., 2020). Moreover, the projected growth of steel production calls for smart and adaptive systems that can adapt to changing socio-technical and environmental circumstances (Vögele et al., 2020). Recent studies have highlighted the need for holistic solutions capable of simultaneously managing energy efficiency, emissions and economic factors, underlining the need for data-based methods in industrial systems (Na et al., 2025).

The increased availability of data has provided opportunities for the use of machine learning approaches in energy analysis and prediction. Data-driven methods have shown the potential to enhance performance prediction, resource allocation, and decision-making in steel production (Boto et al., 2022). The adoption of artificial intelligence technologies in industrial systems has also sped up the development of intelligent manufacturing, allowing for better representation of complex interactions among process variables and energy use (Geng et al., 2023). These advancements have implications for studying the interactions of energy at shorter time frames and operational levels, where conventional methods may not be effective. Previous research has addressed different energy topics in the steel industry, such as decomposition of energy consumption and carbon emissions at the site level and the identification of key factors (Wang & Sun, 2024). Machine learning techniques have also been used to improve energy efficiency through prediction models and proactive management (Zhang et al., 2025). Similarly, predictive models have been created to estimate electricity consumption under various operating scenarios, further showcasing the use of data-driven approaches in industry (Al-shaibani et al., 2023). This has also been applied in industrial processes like milling, where predictive models have been developed to link energy consumption with process variables and product quality (Yu et al., 2021). However, there is still a need for predictive modeling approaches combined with interpretable analysis of critical factors affecting energy consumption in industrial settings. Much of the existing research is focused on predictive performance, with limited emphasis on understanding the significance of different variables. Moreover, there is a lack of focus on the impact of electrical characteristics such as reactive power and power factor on the energy consumption in industrial plants. This is critical for enabling effective strategies to go beyond prediction to support process optimization.

The current research seeks to understand and forecast industrial energy consumption through machine learning and to determine the factors influencing energy demand. Through statistical analysis and the use of interpretable machine learning techniques, the study aims to offer both predictive accuracy and insights into the key factors influencing energy consumption. The results are expected to inform better energy management strategies and the design of efficient and sustainable industrial systems.

## 2. Methodology

### 2.1 Research Design

The present study uses a quantitative, empirical design to investigate industrial energy consumption patterns. A supervised learning approach was used to study the influence of electrical and

operational parameters on energy consumption. This method integrates statistical and machine learning methods to provide predictive modelling and interpretability of the key variables.

## **2.2 Data Source**

The data in this research were sourced from a publicly available dataset that aims to provide real-time electrical and operational features of an industrial steel plant (V E et al., 2021). This dataset contains energy consumption measurements, along with reactive power, power factor and time variables measured at a consistent time step. It offers a comprehensive picture of industrial energy consumption under different load and operating conditions.

## **2.3 Variables and Feature Specification**

In this study, the dependent variable is the energy used in kilowatt-hours (kWh). Lagging and leading reactive power, lagging and leading power factors, and time of day, hour, day and month are independent variables in this analysis. Dummies of load type and day type were also included following the transformation. CO<sub>2</sub> was not included as it is a direct function of energy consumption, which may lead to an overfit model.

## **2.4 Data Preprocessing**

The data were preprocessed to standardise and prepare for analysis. The date-time variable was transformed into ordered temporal variables, such as hour of the day, day of the month, and month. Nominal variables were converted to binary variables using one-hot encoding, allowing them to be used in machine learning algorithms. Variables were checked for missing values; none were found. The data were then organized for a regression-based modeling.

## **2.5 Exploratory Data Analysis**

A descriptive analysis was undertaken to explore the statistical characteristics of the variables and their inter-relationships. Measures of centrality and variability were calculated. The distribution of energy consumption was assessed, and time series analysis was conducted using aggregation methods. Correlation analysis was performed to determine the strength and direction of associations between the predictor variables and energy consumption, to provide an initial assessment of factors influencing energy consumption.

## **2.6 Model Development**

We built three regression models to forecast energy consumption: Linear Regression, Decision Tree Regressor and Random Forest Regressor. Linear Regression was applied as a baseline to capture linear relationships, and Decision Tree and Random Forest Regressors were used to capture nonlinear relationships and interactions between features. The data was split into training and testing sets with an 80:20 ratio to assess the generalization ability of the models.

## **2.7 Model Evaluation**

The performance of the models was evaluated using conventional regression measures, such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination ( $R^2$ ). These measures capture different aspects of model accuracy. All models were compared to determine the optimal way to predict energy consumption.

## **2.8 Feature Importance Analysis**

To rank the features in the prediction task, feature importance analysis was conducted in the Random Forest model. This approach measures the importance of each variable in the prediction of industrial energy consumption.

## **3. Results**

### 3.1 Descriptive Statistics of the Dataset

The data set was comprised of 35,040 observations of electrical, time and production characteristics of an industrial steel production system. Table 1 provides descriptive statistics of the variables used in the study.

**Table 1. Descriptive statistics of study variables (N = 35040)**

| Variable                       | Mean      | Standard Deviation | Minimum | 25th Percentile | Median   | 75th Percentile | Maximum  |
|--------------------------------|-----------|--------------------|---------|-----------------|----------|-----------------|----------|
| Energy Consumption (kWh)       | 27.387    | 33.444             | 0.00    | 3.20            | 4.57     | 51.238          | 157.18   |
| Lagging Reactive Power (kVarh) | 13.035    | 16.306             | 0.00    | 2.30            | 5.00     | 22.640          | 96.91    |
| Leading Reactive Power (kVarh) | 3.871     | 7.424              | 0.00    | 0.00            | 0.00     | 2.090           | 27.76    |
| Lagging Power Factor           | 80.578    | 18.921             | 0.00    | 63.32           | 87.96    | 99.022          | 100.00   |
| Leading Power Factor           | 84.368    | 30.457             | 0.00    | 99.70           | 100.00   | 100.000         | 100.00   |
| Time of Day (seconds)          | 42750.000 | 24940.534          | 0.00    | 21375.00        | 42750.00 | 64125.000       | 85500.00 |
| Hour of Day                    | 11.500    | 6.922              | 0.00    | 5.75            | 11.50    | 17.250          | 23.00    |
| Day of Month                   | 15.721    | 8.796              | 1.00    | 8.00            | 16.00    | 23.000          | 31.00    |
| Month                          | 6.526     | 3.448              | 1.00    | 4.00            | 7.00     | 10.000          | 12.00    |

Energy consumption values were highly variable, with a range of 0 to 157.18 kWh, a mean of 27.39 kWh, and a median of 4.57 kWh, suggesting a strong right skew. The reactive power variables also exhibited significant variability, especially lagging reactive power, due to the variability in load demand across operating conditions.

### 3.2 Distribution of Energy Consumption

The statistical distribution of energy consumption is shown in Figure 1, which is highly right-skewed, with a large number of observations at lower values, and a long tail towards the higher end of the distribution.

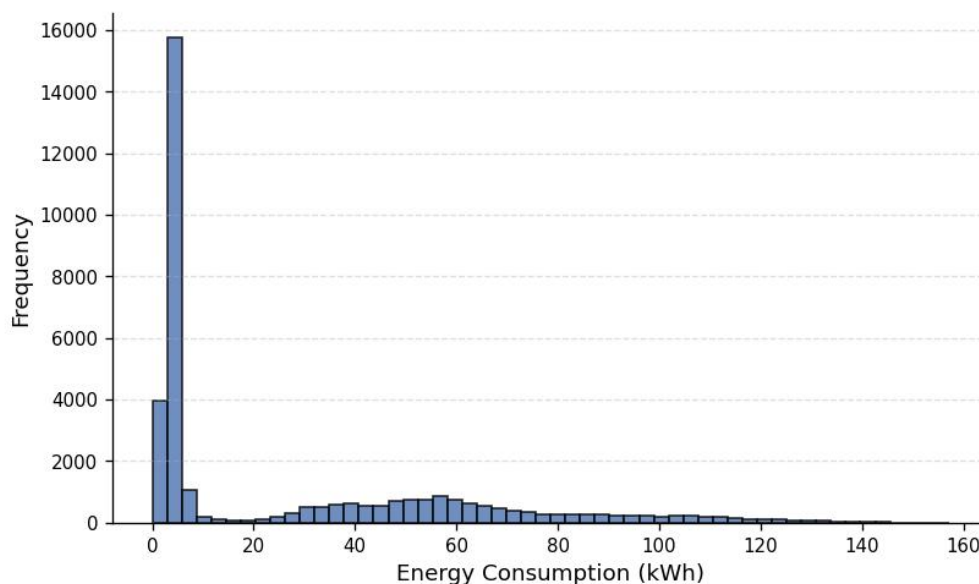


**Figure 1. Distribution of energy consumption (kWh)**

This pattern is a combination of low or no energy consumption and occasional periods of high energy demand, as typically found in industrial production scenarios.

### 3.3 Temporal Variation in Energy Consumption

The temporal behavior of energy usage was analyzed through hourly aggregation. As shown in Figure 2, energy consumption follows a structured daily pattern with clear peaks during specific hours.

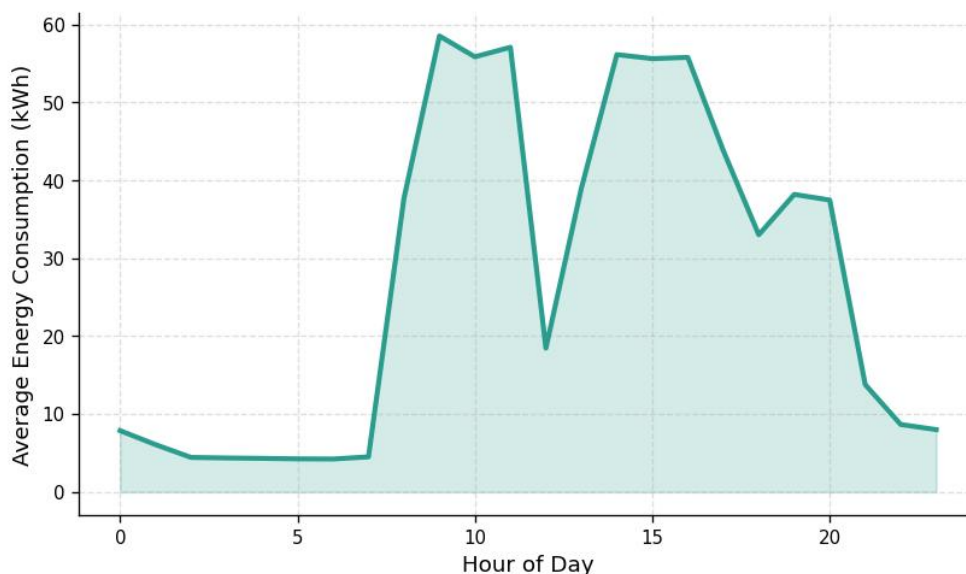


**Figure 2. Average energy consumption by hour of the day**

Peaks are detected in mid-morning and afternoon, with much lower values during the night. This suggests that energy consumption is related to industrial activities.

### 3.4 Correlation Analysis of Predictor Variables

We performed a correlation analysis between the predictors and energy consumption (excluding CO<sub>2</sub> emissions to prevent redundancy). The results are presented in Figure 3.

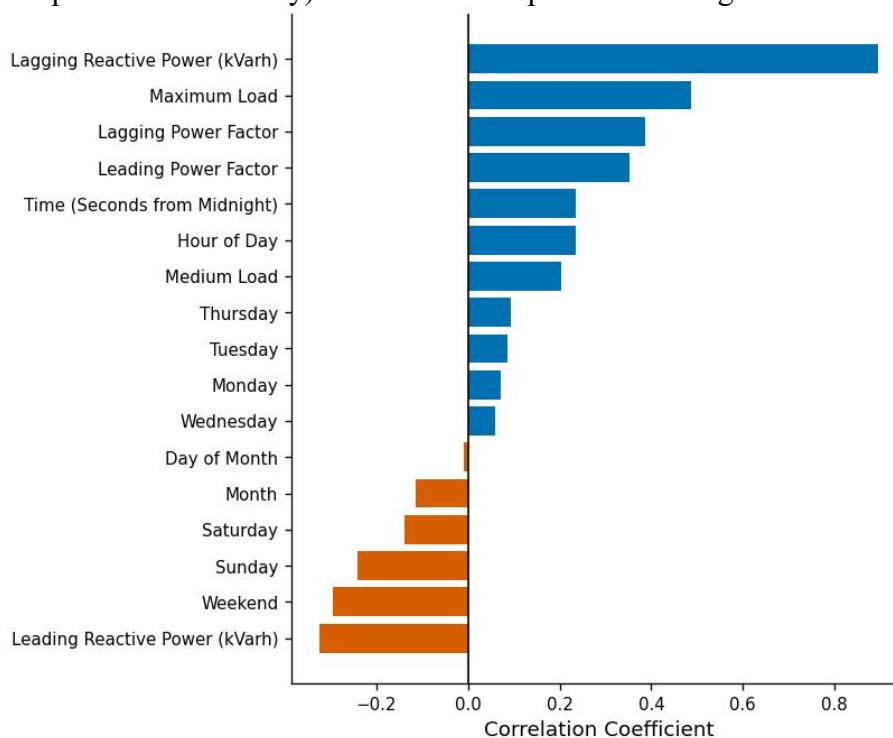


**Figure 3. Correlation of predictor variables with energy consumption**

The highest positive correlation with power consumption was for lagging reactive power, followed by load and power factor. Time-related variables exhibited weaker correlations, implying that the electrical variables are more closely related to energy consumption.

### 3.5 Model Performance Evaluation

The accuracy of three machine learning models was evaluated and is shown in Table 2. Random Forest was the most accurate model with the lowest error values and an  $R^2$  value close to 1. The Decision Tree was also accurate, with the Linear Regression having relatively low accuracy, revealing nonlinear relationships in the data.

**Table 2. Performance comparison of machine learning models**

| Model             | Mean Absolute Error | Root Mean Squared Error | $R^2$ Score |
|-------------------|---------------------|-------------------------|-------------|
| Random Forest     | 0.2917              | 0.8864                  | 0.9993      |
| Decision Tree     | 0.5740              | 1.4978                  | 0.9980      |
| Linear Regression | 7.0405              | 9.6509                  | 0.9181      |

### 3.6 Feature Importance Analysis

The importance of individual predictors was assessed using the feature importance from the Random Forest model. These are shown in Figure 4 and Table 3.

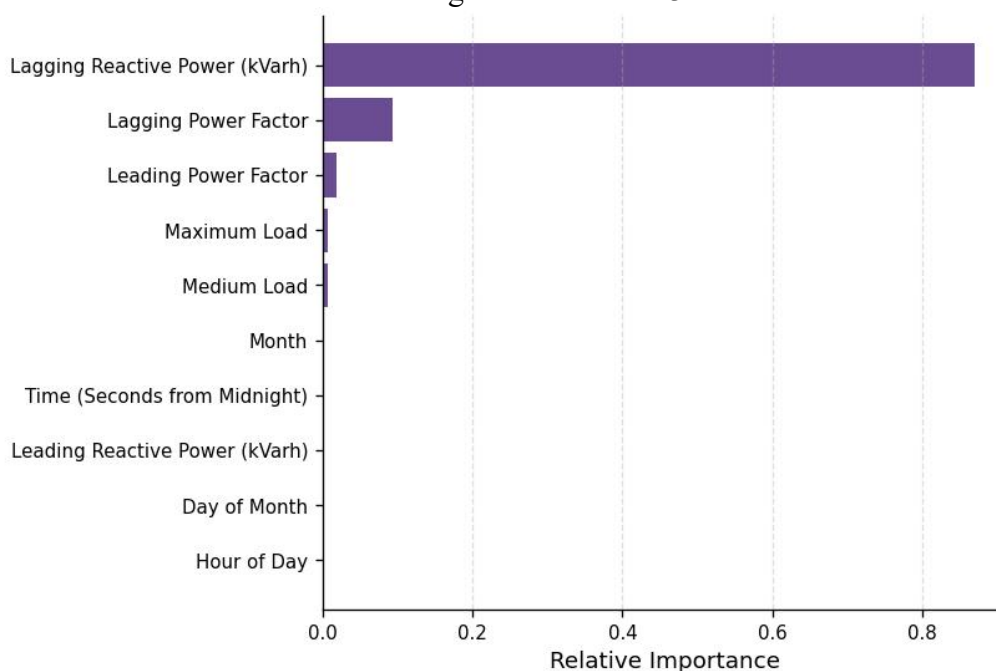


**Figure 4. Key drivers of energy consumption based on feature importance**

**Table 3. Top predictors of energy consumption**

| Predictor Variable             | Relative Importance | Relative Importance (%) |
|--------------------------------|---------------------|-------------------------|
| Lagging Reactive Power (kVarh) | 0.8703              | 87.0332                 |
| Lagging Power Factor           | 0.0931              | 9.3144                  |
| Leading Power Factor           | 0.0192              | 1.9206                  |
| Maximum Load                   | 0.0076              | 0.7599                  |
| Medium Load                    | 0.0061              | 0.6128                  |
| Month                          | 0.0016              | 0.1646                  |
| Time of Day (seconds)          | 0.0007              | 0.0727                  |
| Leading Reactive Power (kVarh) | 0.0007              | 0.0685                  |
| Day of Month                   | 0.0002              | 0.0179                  |
| Hour of Day                    | 0.0002              | 0.0158                  |

The most influential predictor was lagging reactive power, which explained most of the variance. The remaining predictors play a minor role, suggesting that characteristics of electrical load have a significant influence on energy consumption.

### 3.7 Model Validation and Residual Analysis

The accuracy of the Random Forest model was also assessed graphically. Figure 5 shows the graph of the predicted values against actual values.

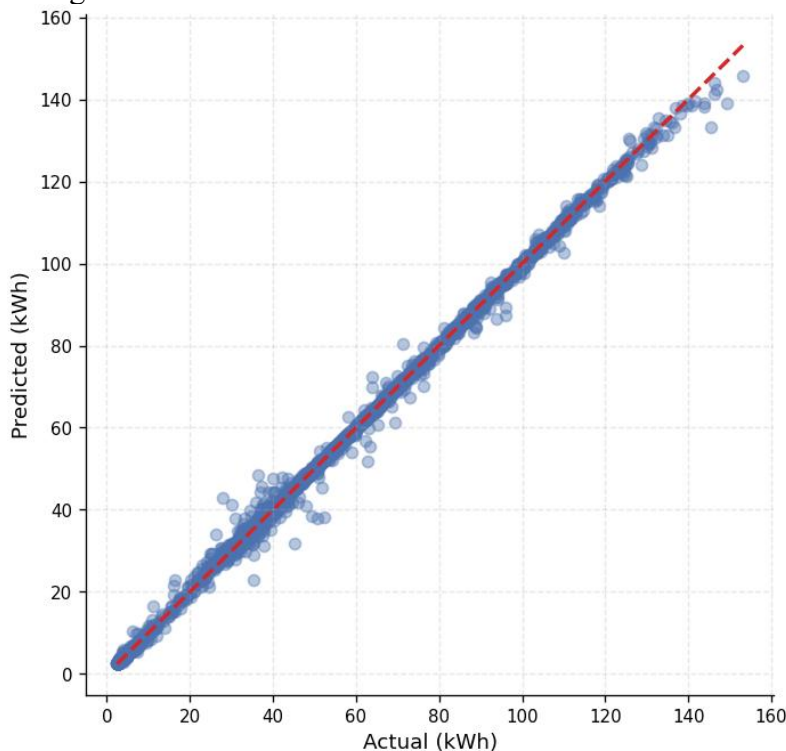


**Figure 5. Actual versus predicted energy consumption using Random Forest**

The observations are closely aligned with the diagonal, showing a close match between actual and predicted values, suggesting the accuracy of the model. Figure 6 displays the residuals (errors) of the predictions.

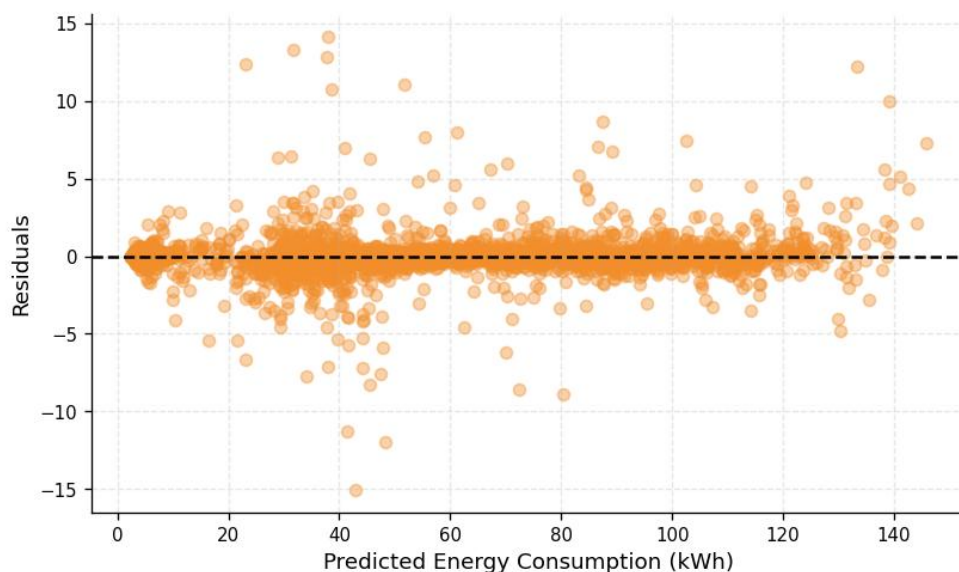


**Figure 6. Residual analysis of the Random Forest model**

The residuals are distributed around zero with no obvious bias, meaning that the model does not have appreciable bias. There is some variability in the middle ranges, but otherwise the model is fairly stable over the energy consumption range.

#### 4. Discussion

Energy demand in the industry had significant variability, with an extremely skewed distribution and distinct periods of low and high energy consumption. The broad variation in energy consumption shows that the operating conditions were not the same throughout the data collection period, due to the non-continuous and process-intensive nature of steel production. The hourly trend also indicated that energy consumption followed a consistent pattern throughout the day, with peaks during production and reduced consumption during non-peak hours. This indicates that energy use was impacted not only by electrical factors but also by the operational schedule of the plant. The results from the machine learning approach showed that nonlinear models outperformed the linear model. The Random Forest model performed the best, followed by Decision Tree, with Linear Regression having relatively higher errors. This suggests that nonlinear relationships between electrical and operating variables influence industrial energy demand. The superior performance of tree-based methods is in line with the nature of industrial systems, in which energy consumption may exhibit significant jumps between different categories of loads and operating points. The feature importance analysis revealed that lagging reactive power is the most significant variable for predicting energy demand, contributing to the model much more than temporal or categorical variables. This is technically significant because reactive power is related to the electrical loads in industrial systems, particularly when there are motors, transformers and inductive loads. Power factor features also predicted consumption, albeit to a lesser degree, suggesting energy consumption is affected by electrical efficiency. Although time-based features contributed to describing operational behavior, their low feature importance indicates that electrical features have a greater influence on consumption than time.

The findings are in line with previous research that demonstrates that energy and environmental performance in iron and steel manufacturing are influenced by process performance and operational efficiency. Gu et al. (2023) highlighted the importance of assessing the energy, environmental, and economic potential in parallel, suggesting that energy prediction can help improve overall industrial performance, as found here. Likewise, Gao et al. (2020) identified technological opportunities for energy and resource saving in steel manufacturing, confirming the usefulness of assessing the drivers of consumption. The ability of machine learning models to accurately predict energy consumption is also in line with other studies on smart factory energy prediction. Sathishkumar et al. (2020) showed the value of data mining techniques for predicting energy consumption, and Carlsson et al. (2020) showed that interpretable machine learning can provide insights into electrical energy consumption in electric arc furnaces. Our results build on this by demonstrating that accurate prediction and interpretability can go hand in hand using model comparison and feature importance.

The role of lagging reactive power is also important in the context of studies of electrical network behaviour. Sayenko et al. (2023) studied reactive power in networks with non-linear and fast-varying loads, reinforcing the view that reactive power is important in industrial electrical loads. The relevance of power characteristics also aligns with Shin et al. (2019), who studied electrical power characteristics of iron and steel plants, and highlighted the economic importance of energy-system design. Our results also support recent research on flexible and real-time industrial energy management. Liu et al. (2025) established the benefits of demand response in steel processes for energy efficiency and environmental improvement, while Iannino et al. (2022) revealed the usefulness of flexible scheduling for flat steel production. While the current study did not include scheduling, the hourly pattern observed for consumption suggests that predictive models will be useful for scheduling and demand management in the future. Furthermore, Tusnin et al. (2024) demonstrated the general applicability of machine learning for load-related patterns in structures, showing that machine learning is also useful for load-related patterns in other engineering applications.

The results are relevant to industrial energy efficiency. Given that lagging reactive power was the most significant predictor, energy efficiency can possibly be improved by monitoring and managing

reactive power. Organizations can use this information to focus on power factor correction, load balancing and equipment diagnostics. The success of Random Forest also implies that machine learning techniques can be used to inform short-term energy predictions, allowing managers to plan for energy peaks and better manage energy resources. In terms of sustainability, better forecasts of energy demand can help to reduce waste and optimise resource utilisation. While this study did not directly model emissions, energy efficiency is directly related to environmental sustainability in energy-intensive sectors. The findings, therefore, give an indication of how predictive analytics can be linked to sustainable manufacturing practices in decision support systems.

A few limitations should be noted. Secondary data were used in the analysis, which restricts control of variable selection, measurement methods and plant characteristics. Critical variables, such as production rate, equipment status, raw material type, shift pattern and maintenance tasks were not included. The extremely high model accuracy may be overfit to the operational patterns in the data, so the model should be tested on other industrial plants before deployment.

Future work should use real-time plant data and other operational factors to enhance generalizability. Future studies could also explore other algorithms, time-series cross-validation, and explainable algorithms (such as SHAP) to improve interpretability. Real-world deployment studies may test the impact of predictive models on production scheduling, peak load management, cost control, and energy-efficiency measures in manufacturing.

## 5. Conclusion

Energy demand projections are crucial for enhancing efficiency and enabling sustainable industrial operations. The study showed that machine learning algorithms can capture energy consumption patterns in an industrial steel manufacturing. Random Forest performed best of the models tested, suggesting it is well-suited to model nonlinear interactions between electrical and process parameters. Lagging reactive power was the most significant predictor, followed by power factor-related features, demonstrating that electrical load properties are important in industrial energy consumption. Time-based features explained daily energy consumption patterns but were less important for prediction accuracy. The results underline the importance of predictive energy models with feature-level interpretation for industrial energy management. This can aid in monitoring, forecasting, power factor adjustment, and planning. But interpretation is limited by the use of secondary data. Future research should use real-time industrial data for model validation and consider other operational factors for better generalizability and applicability.

## References

1. V E, S., Shin, C., & Cho, Y. (2021). Steel Industry Energy Consumption [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C52G8C>.
2. Holappa, L. (2020). A general vision for the reduction of energy consumption and CO2 emissions from the steel industry. *Metals*, 10(9), 1117.
3. Conejo, A. N., Birat, J. P., & Dutta, A. (2020). A review of the current environmental challenges of the steel industry and its value chain. *Journal of environmental management*, 259, 109782.
4. Ahmad, I., Arif, M. S., Cheema, I. I., Thollander, P., & Khan, M. A. (2020). Drivers and barriers for efficient energy management practices in energy-intensive industries: a case-study of iron and steel sector. *Sustainability*, 12(18), 7703.
5. Vögele, S., Rübbelke, D., Govorukha, K., & Grajewski, M. (2020). Socio-technical scenarios for energy-intensive industries: the future of steel production in Germany. *Climatic Change*, 162(4), 1763-1778.
6. Na, H., Yuan, Y., Sun, J., Zhang, L., & Du, T. (2025). Integrative optimization for energy efficiency, CO2 reduction, and economic gains in the iron and steel industry: A holistic approach. *Resources, Conservation and Recycling*, 212, 107992.
7. Boto, F., Murua, M., Gutierrez, T., Casado, S., Carrillo, A., & Arteaga, A. (2022). Data driven performance prediction in steel making. *Metals*, 12(2), 172.

8. Geng, X., Wang, F., Wu, H. H., Wang, S., Wu, G., Gao, J., ... & Mao, X. (2023). Data-driven and artificial intelligence accelerated steel material research and intelligent manufacturing technology. *Materials Genome Engineering Advances*, 1(1), e10.
9. Wang, J., & Sun, W. (2024). Decomposition of the site-level energy consumption and carbon dioxide emissions of the iron and steel industry. *Environmental Science and Pollution Research*, 31(11), 16511-16529.
10. Zhang, X., Yang, L., & Ma, D. (2025). The Importance of Energy Consumption and the Need for Efficiency in the Steel Industry Using Machine Learning. *Journal of Electrical Engineering & Technology*, 20(8), 5097-5112.
11. Al-shaibani, W. T., Babaqi, T., & Alsarori, A. (2023). Power consumption prediction for steel industry. *arXiv preprint arXiv:2307.07597*.
12. Yu, S., Zhao, G., Li, C., Xu, S., & Zheng, Z. (2021). Prediction models for energy consumption and surface quality in stainless steel milling. *The International Journal of Advanced Manufacturing Technology*, 117(11), 3777-3792.
13. Gu, Y., Liu, W., Wang, B., Tian, B., Yang, X., & Pan, C. (2023). Analysis and prediction of energy, environmental and economic potentials in the iron and steel industry of China. *Processes*, 11(12), 3258.
14. Sathishkumar, V. E., Lee, M., Lim, J., Kim, Y., Shin, C., Park, J., & Cho, Y. (2020). An energy consumption prediction model for smart factory using data mining algorithms. *KIPS Transactions on Software and Data Engineering*, 9(5), 153-160.
15. Gao, C. K., Na, H. M., Song, K., Tian, F., Strawa, N., & Du, T. (2020). Technologies-based potential analysis on saving energy and water of China's iron and steel industry. *Science of the total environment*, 699, 134225.
16. Carlsson, L. S., Samuelsson, P. B., & Jönsson, P. G. (2020). Interpretable machine learning—tools to interpret the predictions of a machine learning model predicting the electrical energy consumption of an electric arc furnace. *steel research international*, 91(11), 2000053.
17. Shin, H. K., Cho, J. M., & Lee, E. B. (2019). Electrical power characteristics and economic analysis of distributed generation system using renewable energy: Applied to iron and steel plants. *Sustainability*, 11(22), 6199.
18. Sayenko, Y., Pawelek, R., & Baranenko, T. (2023). Analysis of Reactive Power in Electrical Networks Supplying Nonlinear Fast-Varying Loads. *Energies*, 16(24), 8011.
19. Liu, X., Sun, W., Chen, T., Xu, X., & Huang, T. (2025). Energy and environmental performance of iron and steel industry in real-time demand response: A case of hot rolling process. *Applied Energy*, 389, 125717.
20. Iannino, V., Colla, V., Maddaloni, A., Brandenburger, J., Rajabi, A., Wolff, A., ... & Schirm, C. (2022). A hybrid approach for improving the flexibility of production scheduling in flat steel industry. *Integrated Computer-Aided Engineering*, 29(4), 367-387.
21. Tusnin, A. R., Alekseytsev, A. V., & Tusnina, O. A. (2024). Load identification in steel structural systems using machine learning elements: uniform length loads and point forces. *Buildings*, 14(6), 1711.