Nusantara Journal of Multidisciplinary Informatics

Volume 1 Issue 1
https://doi.org/10.56705/njomi.v1i1.5

Research Article

Predicting the Rule, Not the Load: A Construct-Validity and Validation-Leakage Audit of Steel Industry Load-Type Classification

Nurul Rismayanti[1]
[1] Universitas Muslim Indonesia, Jl. Urip Sumoharjo No.km.5, Panaikang, Kec. Panakkukang, Kota Makassar, Sulawesi Selatan 90231, Indonesia, nurulrismayanti.labfik.umi.ac.id

Correspondence should be addressed to Nurul Rismayanti; nurulrismayanti.labfik@umi.ac.id

Received 15 February 2026 Accepted 29 March 2026; Published 30 April 2026

Abstract:
Introduction: The Steel Industry Energy Consumption dataset has been widely used for machine-learning classification of `Load_Type`. However, the provenance of this target and the effect of temporal structure on classification performance have received limited attention. This study evaluates whether high classification accuracy reflects physical industrial load recognition or reconstruction of predefined temporal rules. Method: The study analyzed 35,040 observations recorded at 15-minute intervals throughout 2018. Daily label sequences were audited and compared with the seasonal and time-of-day schedules documented in the original dataset source. A deterministic rule-based classifier was developed without machine learning. Random Forest was then evaluated using calendar-only, electrical-only, and combined feature sets under random-row, grouped-day, and chronological validation strategies. Results: Only three unique daily `Load_Type` patterns were identified across 365 days, with 303 days following two documented seasonal schedules. The deterministic baseline achieved 90.09% accuracy and correctly reconstructed all Medium and Maximum Load observations. Under chronological validation, calendar and seasonal variables achieved 92.81% accuracy, compared with 66.30% using electrical variables alone. Including exact calendar position produced 99.64% accuracy under random-row splitting but decreased to 71.80% under chronological evaluation. Conclusion: `Load_Type` contains a strong deterministic temporal component, and random-row validation can substantially overestimate temporal generalization. Classification performance should therefore be interpreted according to label provenance, predictor structure, and intended deployment conditions.

Keywords: Construct Validity; Industrial Energy Consumption; Load Classification; Machine Learning; Temporal Validation; Validation Leakage.

1. Introduction

Energy-intensive industries increasingly use machine learning to support energy management and operational efficiency. In steel manufacturing, electricity consumption is influenced by production schedules, operating conditions, and temporal patterns, making data-driven analysis relevant for understanding and predicting energy behavior [1]–[7].

The Steel Industry Energy Consumption Dataset is frequently used for this purpose. It contains 35,040 observations recorded at 15-minute intervals and includes energy consumption, reactive power, power factors, temporal information, and a categorical variable called Load_Type [8]. Although the dataset was originally associated with energy-consumption prediction, several studies have subsequently used Load_Type as a classification target and reported accuracies above 90% using Random Forest, Decision Tree, explainable machine learning, and TinyML approaches [9]–[13].

However, the original dataset publication explicitly defines Light, Medium, and Maximum Load according to time-of-day intervals that vary across seasonal periods [1]. The dataset also provides NSM, representing seconds from midnight, together with other calendar information. This raises an important question: high classification accuracy may reflect reconstruction of the temporal rules used to create the labels rather than recognition of physical industrial load conditions. Such a distinction is relevant to construct validity because predictive performance does not necessarily confirm that the target represents the physical phenomenon implied by its name [14], [15].

The evaluation strategy may further influence the reported performance. Random row-level splitting can place observations from the same date in both training and testing sets. In temporally structured data, this can lead to optimistic performance estimates, particularly when calendar-related predictors are available [16]-[20]. Therefore, evaluation should consider whether the model generalizes to unseen days and future periods rather than only to randomly selected observations.

This study investigates Load_Type classification from this perspective. The analysis reconstructs the documented temporal rules without machine learning, compares the predictive contribution of calendar and electrical variables, and evaluates model performance using random-row, grouped-day, and chronological validation. The study aims to determine whether high Load_Type classification accuracy primarily represents physical load recognition or reconstruction of the labeling structure embedded in the dataset.

2. Method

Dataset and Research Design

This study used the Steel Industry Energy Consumption Dataset from the UCI Machine Learning Repository [8]. The dataset contains 35,040 observations recorded every 15 minutes throughout 2018, corresponding to 96 observations per day for 365 days. The target variable, Load_Type, consists of Light_Load, Medium_Load, and Maximum_Load.

The analysis was designed as a diagnostic audit rather than an algorithm-comparison study. It consisted of four stages: daily label-pattern analysis, deterministic schedule reconstruction, comparison of temporal and electrical feature groups, and evaluation under different train-test strategies.

Label-Pattern Audit

The observations were ordered by date and `NSM`. For each day, the 96 consecutive labels were represented as a daily sequence:

$$L_d = [y_{d,1}, y_{d,2}, \dots, y_{d96}] \tag{1}$$

Daily sequences were compared to identify recurring labeling patterns across the year. Days containing `Light_Load` for all 96 intervals were treated separately as special all-Light days, without assuming their operational cause.

Rule-Based Reconstruction

A deterministic baseline was developed using the seasonal and time-of-day load schedules documented in the original dataset publication [1]. The predicted class was determined only by season and `NSM`:

$$\hat{y}_i^{rule} = f(S_i, NSM_i) \tag{2}$$

No electrical measurements or machine-learning training were used. This experiment measured how much of `Load_Type` could be reconstructed directly from its documented temporal rules.

Feature Groups and Classification

Random Forest was used as a diagnostic classifier with the same configuration across experiments. Predictor variables were divided into controlled groups as shown in **Table 1**.

Table 1. Feature groups used in the experiments

Feature group	Main variables
NSM only	NSM
Calendar basic	NSM, day of week, weekend status
Calendar + season	Calendar basic + seasonal indicator
Calendar + exact date	Calendar basic + day of year
Electrical only	Usage, reactive power, CO_2, power factors
Electrical + calendar	Electrical variables + calendar and season
All + exact date	All predictors + day of year

The electrical-only experiment excluded all temporal predictors. Conversely, calendar-only models excluded electrical measurements. `Day_of_year` was included only as an ablation variable to examine whether exact calendar information could inflate classification performance.

Validation Strategies

Three 80:20 validation strategies were applied. Each experiment contained 28,032 training observations and 7,008 testing observations.

- Random-row split randomly divided individual 15-minute observations without considering their date.
- Grouped-day split assigned all 96 observations from the same date exclusively to either training or testing, preventing same-day overlap.
- Chronological split used the first 292 days, from 1 January to 19 October 2018, for training and the remaining 73 days, from 20 October to 31 December 2018, for testing.

These strategies were used to distinguish random interpolation, generalization to unseen days, and prediction of future periods.

Evaluation

Model performance was assessed using accuracy, balanced accuracy, macro F1, and confusion matrices. Accuracy was calculated as:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\,(y_i - \hat{y}_i) \tag{3}$$

Balanced accuracy and macro F1 were included because the three target classes were not equally distributed.

To quantify sensitivity to validation design, the performance difference between random and chronological evaluation was calculated as:

$$\Delta_{R-C} = Metric_{Random} - Metric_{chronological} \tag{4}$$

A larger value indicates that performance obtained from random splitting transfers poorly to future periods.

3. Result and Discussion

Daily Label Structure

The dataset contains 18,072 `Light_Load`, 9,696 `Medium_Load`, and 7,272 `Maximum_Load` observations. However, the temporal audit showed that only three unique daily label patterns occurred across the entire year.

Table 2. Distribution of Unique Daily Load-Type Patterns

Pattern	Days	Description
A	62	All Light Load

Pattern	Days	Description
B	95	November–February schedule
C	208	March–October schedule

Thus, 303 of 365 days (83.01%) followed one of two recurring seasonal schedules. The remaining 62 days were entirely classified as Light Load. These special all-Light days also showed much lower average energy consumption, 3.65 kWh, compared with 32.24 kWh on regular schedule days.

This result indicates that Load_Type has a strong predetermined temporal structure rather than behaving solely as a physical state derived from electrical measurements.

Rule-Based Reconstruction

The deterministic schedule reconstructed Load_Type with 90.09% accuracy, 93.60% balanced accuracy, and 90.27% macro F1, without machine learning or electrical predictors.

After quality filtering, the common modeling cohort consisted of 31,732 observations. Random Forest performance under the five feature configurations is presented in **Table 3**.

Table 3. Confusion Matrix of the Deterministic Rule-Based Reconstruction

Actual	Light	Medium	Maximum
Light	14,600	1,984	1,488
Medium	0	9,696	0
Maximum	0	0	7,272

All Medium and Maximum Load observations were correctly reconstructed. Overall, 31,568 of 35,040 observations were predicted correctly using only the documented temporal schedule.

This finding is important because previous machine-learning studies using the same target have reported accuracy in a similar range [5]–[9]. Therefore, high classification accuracy alone does not necessarily indicate that a model has learned complex industrial load behavior.

Calendar Versus Electrical Features

The Random Forest experiments produced clear differences between temporal and physical predictors.

Table 4. Classification Accuracy Across Feature Sets and Validation Strategies

Feature set	Random	Grouped-day	Chronological
NSM only	81.02%	79.00%	66.44%
Calendar basic	90.64%	90.71%	71.80%
Calendar + season	**92.84%**	**92.81%**	**92.81%**
Calendar + exact date	**99.64%**	91.47%	71.80%
Electrical only	74.16%	71.12%	**66.30%**
Electrical + calendar	98.40%	97.32%	91.20%
All + exact date	99.22%	97.07%	73.26%

Under chronological validation, `Calendar + season` achieved 92.81% accuracy, while electrical variables alone achieved only 66.30%. The difference of 26.51 percentage points shows that temporal information carries substantially more information about `Load_Type` than the electrical measurements.

Adding electrical variables improved random-split performance, but this advantage was reduced under chronological evaluation. This suggests that some relationships learned from the electrical variables may be specific to the observed period rather than consistently transferable to future dates.

Effect of Validation Strategy

The strongest validation effect appeared when exact calendar position was included. Under random-row splitting, all 365 dates appeared in both training and testing sets.

With `day_of_year`, accuracy reached **99.64%** under random splitting, but decreased to **91.47%** with grouped-day validation and **71.80%** with chronological validation.

The performance difference was:

$$\Delta_{RC} = 99.64\% - 71.80\% = 27.84\%$$

In contrast, the `Calendar + season` model remained highly stable at **92.84%, 92.81%, and 92.81%** across random, grouped-day, and chronological evaluation.

This contrast suggests that the seasonal schedule generalizes across dates, whereas the near-perfect performance obtained with exact-date information is strongly dependent on the validation design. Random row-level splitting can therefore overestimate performance when temporally related observations from the same dates occur in both training and testing sets [3], [12].

Interpretation and Implications

The combined results indicate that Load_Type is better interpreted as a hybrid operational label with a strong temporal scheduling component rather than as a purely physical industrial load state. This does not make the dataset unsuitable for machine learning. It remains useful for schedule classification, exception detection, and energy analytics. However, researchers should distinguish between reconstructing predefined scheduling rules and predicting actual physical operating conditions.

For future studies, deterministic baselines should be considered whenever the target is partly defined by known operational rules. Grouped-day or chronological validation should also be preferred when the intended application involves unseen days or future periods. Overall, the results demonstrate that understanding how a target is constructed is essential for correctly interpreting machine-learning performance.

Recent classification studies further demonstrate that model performance is strongly influenced by feature representation, learning strategy, preprocessing, class distribution, and evaluation design. Studies involving Azis and collaborators have shown substantial variations in classification performance across segmentation strategies, supervised and semi-supervised learning settings, class-balancing techniques, and deep-learning architectures [21]–[25]. These findings reinforce the need to interpret performance metrics in relation to the complete experimental design rather than model accuracy alone.

Similar considerations are evident in energy and data-science applications. Studies published in the Indonesian Journal of Data and Science have examined machine-learning approaches for energy forecasting and power-consumption optimization, while recent work using repeated nested cross-validation and independent holdout testing emphasizes the importance of rigorous validation when estimating model generalization [26]–[28]. Moreover, classification studies published in the International Journal of Artificial Intelligence in Medical Issues demonstrate that ensemble strategies and data-resampling procedures can substantially affect reported predictive performance [29], [30]. These studies support the present finding that high classification accuracy should be interpreted together with predictor construction, data preprocessing, and validation strategy.

4. Conclusion

This study examined the construct validity and validation sensitivity of `Load_Type` classification in the Steel Industry Energy Consumption dataset. The analysis showed that only three unique daily label patterns occurred across the year, with 303 of 365 days following two recurring seasonal schedules.

A deterministic rule based only on season and time of day achieved 90.09% accuracy without machine learning or electrical measurements. Furthermore, calendar and seasonal variables achieved 92.81% chronological accuracy, substantially higher than the 66.30% obtained using electrical variables alone.

The validation strategy also strongly influenced performance. Including exact calendar information produced 99.64% accuracy under random-row splitting, but only 71.80% under chronological validation. This indicates that random splitting can substantially overestimate temporal generalization when observations contain strong date-related dependencies.

These findings suggest that Load_Type contains a strong temporal scheduling component and should not be interpreted solely as a physical industrial load state. Future studies should distinguish between schedule reconstruction and physical load-state prediction, use appropriate temporal validation, and include deterministic baselines when known labeling rules exist.

Overall, high classification accuracy should be interpreted together with label provenance, predictor structure, and validation design to ensure that model performance reflects the intended research problem.

References

[1] V. E. Sathishkumar, C. Shin, and Y. Cho, "Efficient energy consumption prediction model for a data analytic-enabled industry building in a smart city," Building Research & Information, vol. 49, no. 1, pp. 127–143, 2021, doi: https://doi.org/10.1080/09613218.2020.1809983.

[2] J. Walther and M. Weigold, "A systematic review on predicting and forecasting the electrical energy consumption in the manufacturing industry," Energies, vol. 14, no. 4, Art. no. 968, 2021, doi: https://doi.org/10.3390/en14040968.

[3] F. Gu and Y. Zhao, "Comparative evaluation of several models for forecasting hourly electricity use in a steel plant," Scientific Reports, vol. 16, Art. no. 13123, 2026, doi: https://doi.org/10.1038/s41598-026-43868-z.

[4] D. A. C. Narciso and F. G. Martins, "Application of machine learning tools for energy efficiency in industry: A review," Energy Reports, vol. 6, pp. 1181–1199, 2020, doi: https://doi.org/10.1016/j.egyr.2020.04.035.

[5] V. J. Mawson and B. R. Hughes, "Deep learning techniques for energy forecasting and condition monitoring in the manufacturing sector," Energy and Buildings, vol. 217, Art. no. 109966, 2020, doi: https://doi.org/10.1016/j.enbuild.2020.109966.

[6] Y. Hu and Y. Man, "Energy consumption and carbon emissions forecasting for industrial processes: Status, challenges and perspectives," Renewable and Sustainable Energy Reviews, vol. 182, Art. no. 113405, 2023, doi: https://doi.org/10.1016/j.rser.2023.113405.

[7] S. Ma, Y. Zhang, J. Lv, Y. Ge, H. Yang, and L. Li, "Big data driven predictive production planning for energy-intensive manufacturing industries," Energy, vol. 211, Art. no. 118320, 2020, doi: https://doi.org/10.1016/j.energy.2020.118320.

[8] V. E. Sathishkumar, C. Shin, and Y. Cho, "Steel Industry Energy Consumption," UCI Machine Learning Repository, 2021, doi: https://doi.org/10.24432/C52G8C.

[9] J. J. Purnama and S. Rahayu, "Klasifikasi konsumsi energi industri baja menggunakan teknik data mining," Jurnal Teknoinfo, vol. 16, no. 2, pp. 395–407, 2022, doi: https://doi.org/10.33365/jti.v16i2.1984.

[10] M. Chan, "Evaluasi efektivitas algoritma klasifikasi beban penggunaan listrik pada mesin pabrik baja," Jurnal Komputer dan Informatika, vol. 20, no. 1, pp. 1–8, 2025.

[11] G. Tzionis, G. Kougka, I. Gialampoukidis, S. Vrochidis, I. Kompatsiaris, and M. Vlachopoulou, "Enhancing robustness in feature importance methods with NAFIC and CESHAP for improved interpretability," Applied

Artificial Intelligence, vol. 39, no. 1, Art. no. 2515062, 2025, doi: https://doi.org/10.1080/08839514.2025.2515062.

[12] M. Muthusamy, A. Sureshkumar, P. Shanmugam, P. R. Kaveri, M. Y. N. Mohamed, and S. Sridhar, “TinyML with CTGAN based smart industry power load usage prediction with original and synthetic data visualization towards Industry 5.0,” Scientific Reports, vol. 15, Art. no. 41712, 2025, doi: https://doi.org/10.1038/s41598-025-25678-x.

[13] G. Tzionis, G. Kougka, I. Gialampoukidis, S. Vrochidis, I. Kompatsiaris, and M. Vlachopoulou, “Improving LIME stability via density-awareness: Evaluation and comparison of AKDE-LIME,” Applied Artificial Intelligence, vol. 40, no. 1, Art. no. 2640686, 2026, doi: https://doi.org/10.1080/08839514.2026.2640686.

[14] A. Z. Jacobs and H. Wallach, “Measurement and fairness,” in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’21), 2021, pp. 375–385, doi: https://doi.org/10.1145/3442188.3445901.

[15] B. Hutchinson, N. Rostamzadeh, C. Greer, K. Heller, and V. Prabhakaran, “Evaluation gaps in machine learning practice,” in Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’22), 2022, pp. 1859–1876, doi: https://doi.org/10.1145/3531146.3533233.

[16] S. Kapoor and A. Narayanan, “Leakage and the reproducibility crisis in machine-learning-based science,” Patterns, vol. 4, no. 9, Art. no. 100804, 2023, doi: https://doi.org/10.1016/j.patter.2023.100804.

[17] C. Bergmeir, R. J. Hyndman, and B. Koo, “A note on the validity of cross-validation for evaluating autoregressive time series prediction,” Computational Statistics & Data Analysis, vol. 120, pp. 70–83, 2018, doi: https://doi.org/10.1016/j.csda.2017.11.003.

[18] V. Cerqueira, L. Torgo, and I. Mozetič, “Evaluating time series forecasting models: An empirical study on performance estimation methods,” Machine Learning, vol. 109, pp. 1997–2028, 2020, doi: https://doi.org/10.1007/s10994-020-05910-7.

[19] D. R. Roberts et al., “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure,” Ecography, vol. 40, no. 8, pp. 913–929, 2017, doi: https://doi.org/10.1111/ecog.02881.

[20] H. Hewamalage, K. Ackermann, and C. Bergmeir, “Forecast evaluation for data scientists: Common pitfalls and best practices,” Data Mining and Knowledge Discovery, vol. 37, no. 2, pp. 788–832, 2023, doi: https://doi.org/10.1007/s10618-022-00894-5.

[21] H. Azis, Nirmala, L. Syafie, Herman, F. Fattah, and T. Hasanuddin, “Unveiling algorithm classification excellence: Exploring Calendula and Coreopsis flower datasets with varied segmentation techniques,” in *2024 18th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 2024, pp. 1–7, doi: https://doi.org/10.1109/IMCOM60618.2024.10418246

[22] Purnawansyah, A. P. Wibawa, T. Widiyaningtyas, Haviluddin, H. Darwis, and H. Azis, “An in-depth exploration of supervised and semi-supervised learning on face recognition,” *Open Computer Science*, vol. 15, no. 1, Art. no. 20250029, 2025, doi: https://doi.org/10.1515/comp-2025-0029.

[23] A. Latip, H. Azis, H. Himawan, and D. A. Kurnia, “SMOTE technique utilization in cirrhosis classification: A comparison of Gradient Boosting and XGBoost,” *International Journal of Informatics and Computing*, vol. 1, no. 1, pp. 34–40, 2025.

[24] Harlinda, A. Rendi, H. Azis, D. Indra, L. N. Hayati, and N. Kurniati, “Classification of Cia-Cia letters using MobileNetV2 and CNN methods,” in *2025 19th International Conference on Ubiquitous Information Management and Communication (IMCOM),* 2025, pp. 1–6, doi: https://doi.org/10.1109/IMCOM64595.2025.10857478.

[25] H. Azis, Y. Salim, and Y. Puspitasari, “Deep learning approaches for soft tissue classification of cephalometric images in orthodontics,” *JOIV: International Journal on Informatics Visualization*, vol. 10, no. 3, pp. 988–995, 2026, doi: https://doi.org/10.62527/joiv.10.3.3885.

[26] L. O. Lamidi, A. M. Oyelakin, and M. B. Akinbi, “A survey on machine learning techniques for the prediction of solar power production,” *Indonesian Journal of Data and Science*, vol. 5, no. 2, pp. 109–114, 2024, doi: https://doi.org/10.56705/ijodas.v5i2.130.

[27] R. K. Hondro, H. S. Ginting, P. J. N. Simanjuntak, H. T. Silalahi, and Sarwandi, “Use of machine learning in power consumption optimization of computing devices,” *Indonesian Journal of Data and Science*, vol. 6, no. 1, pp. 103–111, 2025, doi: https://doi.org/10.56705/ijodas.v6i1.231.

[28] E. H. Saputra, I. A. E. Zaeni, D. D. Prasetya, A. M. Zain, W. Antonius, and I. M. Wirawan, “Comparative evaluation of machine learning models for heavy crude oil viscosity prediction using repeated nested cross-validation and independent holdout testing,” *Indonesian Journal of Data and Science*, vol. 7, no. 2, 2026, doi: https://doi.org/10.56705/ijodas.v7i2.455.

[29] Ilham, “Enhancing cardiovascular disease prediction accuracy through an ensemble machine learning approach,” *International Journal of Artificial Intelligence in Medical Issues*, vol. 2, no. 2, pp. 95–103, 2024, doi: https://doi.org/10.56705/ijaimi.v2i2.157.

[30] N. El Furqany, “Optimizing air quality index classification using multiple machine learning models and oversampling techniques,” *International Journal of Artificial Intelligence in Medical Issues*, vol. 3, no. 2, pp. 127–141, 2025, doi: https://doi.org/10.56705/ijaimi.v3i2.322.