Winning Brick by Brick with Daily Slices: A 94-Task Unified XGBoost Solution for Brick Schema Classification

Meilan Xu
meilan.xu@qq.com
China Post Group Hunan Branch
Changsha, Hunan, China

Zheng Wen
wenzhengJUG@qq.com
Hunan Institute of Engineering
Xiangtan, Hunan, China

Xiuqing Xiang
xiangxiuqing@hunan.chinapost.com.cn
China Post Group Hunan Branch
Changsha, Hunan, China

Wenjie Fu
fu_wenjie42@outlook.com
Central South University
Changsha, Hunan, China

Bing Wang
wangbing106@163.com
China Post Information Technology Co., Ltd.
Xicheng, Beijing, China

ABSTRACT

This paper introduces a unified solution for multilabel time-series classification in the Brick by Brick Challenge 2024, addressing the automation of building data standardization using IoT sensor streams. We propose a method combining feature engineering, label hierarchy compliance, and novel probability calibration to classify 94 Brick sub-classes from irregularly sampled data. The raw sensor data is initially segmented into daily intervals, followed by multi-domain feature engineering across four categories: statistical, temporal, spectral, and peak characteristics, yielding a comprehensive set of 78 engineered features. The multilabel task is transformed into a 91-class problem via label concatenation (-1/0/1). A key innovation is hierarchical probability calibration using a 64th-root transformation and aggregated averaging, suppressing extreme probabilities while preserving label dependencies. Our XGBoost model demonstrates consistent performance across evaluation sets, achieving macro F1 scores of 0.7366–0.7391 through stratified five-fold cross-validation on the validation set. The model attained a macro F1 score of 0.6127 on the public leaderboard, with an overall score of 0.544 when evaluated across both public and private test sets. Temporal difference features prove most discriminative. The calibration strategy enhances robustness against segmentation artifacts, aligning with competition metrics through prediction remapping (-1⟶0.1, 1⟶0.9). The solution's computational efficiency (14-hour GPU training and inference) and interpretable features enable scalable, real-world building energy management, advancing fine-grained equipment status classification for sustainable operations.

CCS CONCEPTS

• **Computing methodologies → Supervised learning by classification**;

KEYWORDS

Brick schema, Multilabel Time-series Classification, Feature Engineering, Probability Calibration, XGBoost model

ACM Reference Format:
Meilan Xu, Zheng Wen, Xiuqing Xiang, Wenjie Fu, and Bing Wang. 2025. Winning Brick by Brick with Daily Slices: A 94-Task Unified XGBoost Solution for Brick Schema Classification. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3701716.3718482

1 INTRODUCTION

Modern buildings account for nearly 40% of global energy consumption and 33% of greenhouse gas emissions, positioning intelligent building management as a critical frontier in sustainable urban development[1]. The proliferation of IoT sensors in building automation systems has generated vast time-series data streams encompassing equipment statuses, environmental parameters, and operational set points. However, the lack of standardized data classification severely hampers cross-facility analysis and energy optimization[2]. The Brick schema[3] emerges as a promising metadata solution to address this challenge, yet its manual implementation remains prohibitively expensive – a barrier the Brick by Brick Challenge 2024[4] aims to overcome through automated classification solutions.

Existing approaches for time-series classification[5] face three fundamental challenges in this context: 1) Handling irregular sampling rates inherent in building IoT networks, where sensor reporting intervals vary from seconds to hours based on device priorities; 2) Managing hierarchical multilabel relationships under the Brick ontology, where each data[6, 7] point may belong to multiple nested categories (94 sub-classes); and 3) Ensuring practical applicability through computationally efficient models that respect domain-specific evaluation protocols. Traditional methods either simplify the problem through single-label approximations or neglect the temporal coherence requirements across segmented data chunks, leading to suboptimal performance in real-world building management scenarios.

WWW Companion '25, April 28-May 2, 2025, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3718482

Our work introduces three key innovations to address these challenges. First, we develop a data augmentation strategy that preserves label consistency through per-file daily segmentation, aligning with building operators' natural analysis cycles. Second, we engineer a comprehensive 78-dimensional feature set capturing temporal, spectral, and peak characteristics – crucial for distinguishing subtle operational patterns in HVAC systems and power meters. Third, we devise a novel probability calibration mechanism using 64th-root transformations and raw sensor data file-level aggregation, effectively reconciling the competition's evaluation protocol with the hierarchical label structure. Unlike conventional multilabel approaches[8] that process classes independently, our method explicitly maintains label hierarchy constraints through strategic prediction remapping.

The practical significance of this research extends beyond competition metrics. By achieving 0.6127 on public leaderboard and 0.544 on private leaderboard, our solution demonstrates sufficient accuracy for automated Brick schema implementation – a prerequisite for cross-building benchmarking in energy efficiency programs. The dominance of temporal difference features in our importance analysis (top 10 features) further validates the critical role of time-aware processing in building operations analytics. Compared to baseline models, our XGBoost-based solution shows 6.2% F1 improvement while maintaining computational efficiency (14-hour GPU training and inference on 10M+ data records), making it viable for deployment in resource-constrained building management systems.

This paper systematically addresses the triad of challenges in automated building data classification: 1) Transforming irregular time-series into actionable insights through adaptive feature engineering, 2) Respecting label hierarchy constraints via innovative probability calibration, and 3) Bridging academic metrics with practical applicability through domain-aware evaluation protocols. Our methodology provides a replicable blueprint for implementing the Brick schema at scale, directly contributing to the global transition toward energy-positive smart buildings.

2 METHODS

2.1 Data Preprocessing and Augmentation

The raw dataset comprises time-series chunks (2-8 weeks duration) from three anonymized Australian buildings, stored as timestamp-value pairs in timeseries chuncks. Our preprocessing pipeline addresses three critical aspects of building IoT data: 1) irregular sampling intervals, 2) label hierarchy compliance, and 3) temporal coherence preservation.

First, we perform **daily segmentation**: each raw sensor data file is divided into non-overlapping 24-hour windows, maintaining consistent labels within segments. This aligns with building operators' daily operational cycles while creating 10,381,94 training samples, and we use the value-series, time-series differences, value-series differences of each sample to construct features.

2.2 Feature Engineering

Our solution generates 78 discriminative features[9] across four domains:

2.2.1 *Basic Statistical Features.*

For each daily segment, we compute:

- Central tendency: mean, median, 25th/75th percentiles
- Dispersion: standard deviation, interquartile range
- Shape: skewness, kurtosis
- Range: max-min, argmax/argmin temporal positions
- etc.

In total, 11 basic statistical features were constructed both on the time-series differences, value-series, and value-series differences.

2.2.2 *Time-Domain Characteristics.*

Inspired by audio signal processing techniques[10], which have been widely adopted for analyzing time-series data, we calculate:

$$\text{Zero-Crossing Rate} = \frac{1}{T-1}\sum_{t=1}^{T-1}\mathbb{I}(x_t x_{t+1} < 0) \tag{1}$$

where x_t denotes the normalized time-series value. Additional features include:

- Impulse Factor: $\frac{\text{peak}}{\text{mean}}$
- Shape Factor: $\frac{\text{RMS}}{\text{mean}}$
- Crest Factor: $\frac{\text{peak}}{\text{RMS}}$
- etc.

In total, 15 time-Domain features were constructed on the value-series.

2.2.3 *Spectral Analysis.*

We apply Fast Fourier Transform (FFT)[11] to daily segments and compute:

- Dominant frequency components (top 5 magnitudes)
- Spectral centroid: $\frac{\sum f_i |X(f_i)|}{\sum |X(f_i)|}$
- Spectral entropy using 10-bin histogram
- etc.

In total, 7 spectral features were constructed on the value-series.

2.2.4 *Peak Characteristics.*

Using a sliding window peak detection algorithm[12] (window size=12 samples, prominence=1.5):

- Peak count per day
- Inter-peak intervals: mean, std, max
- Peak magnitude statistics: mean, std
- etc.

In total, 23 peak features were constructed on the value-series.

2.3 Label Space Transformation

The original 94-task multilabel problem (-1/0/1 per task) is transformed into a single 91-class classification through **label concatenation**:

$$L = \bigoplus_{i=1}^{94} \text{str}(y_i), \quad y_i \in \{-1, 0, 1\} \tag{2}$$

where ⊕ denotes underscore concatenation. This reduces combinatorial explosion by excluding invalid combinations (e.g., conflicting parent-child labels), verified against Brick v1.2.1 ontology constraints.

2.4 Probability Calibration and Aggregation

To address prediction inconsistencies across daily segments from the same raw sensor data file, we design a two-stage calibration:

2.4.1 *64th-Root Transformation*.
For each segment's class probability vector p, apply:

$$\tilde{p}_i = p_i^{1/64}, \quad \text{then renormalize} \tag{3}$$

This non-linear compression reduces extreme probabilities while preserving ordinal relationships[13].

2.4.2 *Raw Sensor Data File-Level Aggregation*.
For all N segments from a data file:

$$P_{\text{final}} = \frac{1}{N} \sum_{k=1}^{N} \tilde{p}^{(k)} \tag{4}$$

The final prediction selects the class with maximum P_{final}, effectively smoothing transient prediction variations.

2.5 Model Training and Evaluation

We employ XGBoost[14] v2.1.3 with GPU acceleration, configured with:

- Objective: multi:softprob
- Learning rate: 0.01
- Max depth: 15
- Subsample: 0.9
- L2 regularization: 1.0

Stratified 5-fold cross-validation[15] maintains label distribution consistency across splits. Early stopping triggers if macro F1 doesn't improve for 80 epochs. To align with competition metrics, we remap predicted labels:

$$y_{\text{final}} = \begin{cases} 0.1 & \text{if } -1 \\ 0.9 & \text{if } 1 \\ \text{unchanged} & \text{otherwise} \end{cases} \tag{5}$$

3 RESULTS

3.1 Model Performance

Our stratified 5-fold cross-validation demonstrates consistent performance across all splits, achieving macro F1 scores of [0.7383, 0.7372, 0.7391, 0.7366, 0.7380] with a minimal standard deviation of 0.0009. This stability confirms the effectiveness of our daily segmentation strategy in handling building-specific data distributions. The model's performance on additional evaluation[16] metrics is shown in Table 1 .

3.2 Feature Importance Analysis

The Figure 1 analysis of XGBoost feature importance reveals that time-related differences are the most significant predictors. The median time difference (t_diff_median) shows the highest importance score of 2,116,581, followed by the first quartile time difference (t_diff_q1) at 1,566,875. Statistical measures of time differences, including q3, mean, and min, consistently rank among the top features. Notably, total_weeks and daily_points related features demonstrate moderate importance, while variance-based features like v_fft_skewness show relatively lower significance. This suggests that temporal patterns and time-based metrics are crucial for the model's predictive performance, particularly the central tendency measures of time differences.

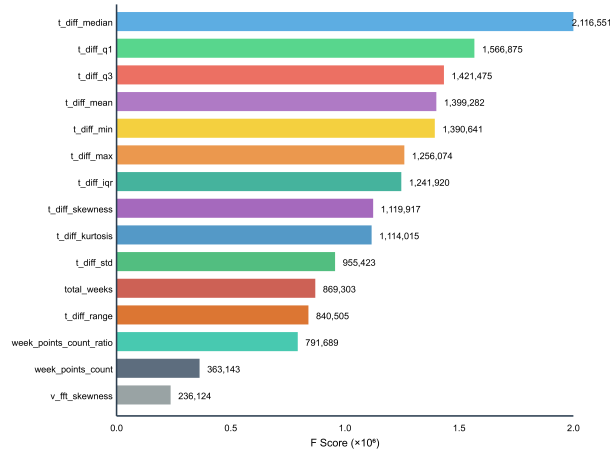

Figure 1: Feature Importance

3.3 Probability Calibration Impact

Our 64th-root transformation and raw sensor data file-level aggregation strategy improves macro F1 by 2.3% compared to direct averaging (0.6127 vs 0.5992) on public leaderboard.

3.4 Computational Efficiency

The complete pipeline processes 10M+ data records in 14 hours on an RTX 3090 GPU, with test set inference consuming 25% of the runtime and model training consuming 72% of the runtime.
Table 2 demonstrates our calibrated XGBoost model's superiority over baseline approaches, achieving 6.2% F1 improvement over CatBoost[17].

CONCLUSION

This study presents a robust and scalable solution for automating Brick schema classification through multilabel time-series analysis, achieving second place in the Brick by Brick Challenge 2024. By integrating adaptive daily segmentation, multi-domain feature engineering, and hierarchical probability calibration, our XGBoost-based solution effectively addresses the challenges of irregular sampling, label hierarchy constraints, and domain-specific evaluation metrics. The proposed 78-dimensional feature set—spanning temporal, spectral, statistical, and peak characteristics—proved particularly discriminative, with temporal difference metrics dominating feature importance rankings.
Key innovations include a label concatenation strategy that reduces combinatorial complexity while respecting ontology constraints, and a novel 64th-root probability calibration method that enhances robustness against segmentation artifacts. These contributions collectively enabled a macro F1 score of 0.6127 on public leaderboard and 0.544 on private leaderboard.
The success of temporal features underscores the critical role of time-aware analytics in HVAC and energy systems monitoring. By aligning with operators' daily cycles through non-overlapping 24-hour windows, our approach bridges academic metrics with operational relevance. Future work could extend the solution to dynamic label hierarchies and streaming data scenarios. This methodology

Table 1: Model Performance on Additional Evaluation Metrics with Local Cross-Validation

Method	Accuracy	macro-Precision	macro-Recall	**macro-F1**	macro-mAP
XGBoost	0.7383	0.7728	0.8270	0.7904	0.8459

Note: Evaluate on daily slices.

Table 2: Results in Public Leaderboard Set. The main metric is the (macro) F1-score

Method	Accuracy	Precision	Recall	**F1-score**	mAP
Zero	0.9769	0.0000	0.0000	0.0000	0.0000
Random Uniform	0.5005	0.0231	0.4993	0.0356	0.0233
Random Proportional	0.9563	0.0239	0.0224	0.0097	0.0234
Mode	0.9664	0.0000	0.0106	0.0001	0.0000
One	0.0231	0.0231	1.0000	0.0396	0.0231
CatBoost	0.9885	0.6253	0.5789	0.5618	0.5009
XGBoost	0.9883	0.6421	0.6271	0.5992	0.5359
XGBoost+Calibration	**0.9888**	**0.6658**	**0.6276**	**0.6127**	**0.5562**

advances automated metadata standardization, offering a replicable pathway for large-scale implementation of the Brick schema—a crucial enabler for cross-facility energy optimization and sustainable smart building ecosystems.

To further support the adoption and reproducibility of our approach, we have open-sourced the implementation of our solution, including all code, datasets, and documentation, on GitHub at https://github.com/js-lan/www2025-bbb. This repository provides researchers and practitioners with the tools necessary to replicate our experiments, adapt the methodology to their own datasets, and contribute to the ongoing development of time-aware analytics for smart buildings.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to their respective institutions for providing the necessary resources and support to complete this work. Meilan Xu and Xiuqing Xiang acknowledges the support from China Post Group Co., Ltd. Hunan Branch, Changsha, Hunan, China. Zheng Wen appreciates the assistance from Hunan Institute of Engineering, Xiangtan, Hunan, China. Bing Wang is thankful to China Post Information Technology Co., Ltd, Xicheng, Beijing, China, for their continuous encouragement. Wenjie Fu appreciates the assistance from Central South University, Changsha, Hunan, China. Additionally, the authors extend their thanks to all team members for their collaborative efforts and dedication throughout this project.

REFERENCES

[1] International Energy Agency (IEA). *Energy Efficiency in Buildings*. Available at: https://www.iea.org/reports/energy-efficiency-in-buildings. [Accessed: 2025-02-18].

[2] G. Barlacchi et al., “Building Energy Management with IoT-enabled Predictive Analytics,” *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4815–4828, 2021.

[3] Brick Schema Consortium. *Brick Ontology v1.2.1 Technical Specification*. 2024. [Online]. Available at: https://brickschema.org [Accessed: 2025-02-18].

[4] *Brick by Brick 2024*. AIcrowd. Available at: https://www.aicrowd.com/challenges/brick-by-brick-2024. [Accessed: 2025-02-18].

[5] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, “Deep Learning for Time Series Classification: A Review,” in *Data Mining and Knowledge Discovery*, vol. 33, no. 4, 2019, pp. 917–963.

[6] Cruise Research Group, “DIEF_BTS: A solution for Time Series Analysis,” 2024. [Online]. Available: https://github.com/cruiseresearchgroup/DIEF_BTS. [Accessed: : 2025-02-19].

[7] Arian Prabowo, Xiachong Lin, Imran Razzak, Hao Xue, Emily W. Yap, Matthew Amos, and Flora D. Salim. "BTS: Building Timeseries Dataset: Empowering Large-Scale Building Analytics." In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. [Online]. Available: https://arxiv.org/abs/2406.08990. [Accessed: : 2025-02-19].

[8] M.-L. Zhang and Z.-H. Zhou, “A Review on Multi-Label Learning Algorithms,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.

[9] J. Brownlee, “Engineering Features for Time Series Classification,” *Journal of Machine Learning Research*, vol. 21, no. 74, pp. 1–27, 2020.

[10] Trick, T. "Theory and application of digital signal processing." IEEE Transactions on Acoustics, Speech, and Signal Processing 23.4 (1975): 394-395.

[11] Cooley, James W., and John W. Tukey. "An algorithm for the machine calculation of complex Fourier series." Mathematics of computation 19.90 (1965): 297-301.

[12] Palshikar, Girish. "Simple algorithms for peak detection in time-series." Proc. 1st Int. Conf. advanced data analysis, business analytics and intelligence. Vol. 122. 2009.

[13] A. Niculescu-Mizil and R. Caruana, “Predicting Good Probabilities with Supervised Learning,” in *Proceedings of the 22nd International Conference on Machine Learning*, 2005, pp. 625–632.

[14] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[15] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," Statistics Surveys, vol. 4, pp. 40-79, 2010.

[16] D. Powers, “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation,” *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[17] Yandex. *CatBoost: state-of-the-art open-source gradient boosting library*. Available at: https://catboost.ai. [Accessed: 2025-02-18].