© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1719689

# A Data Analytics Perspective on Energy Benchmarking in Campus Buildings: A Review of Recent Research

DR. SANGEETA JOSHI[1], LALIT KUMAR JOSHI[2]
[1]*Associate Professor, P.G Department of Computer Science, Mata Gujri College, Fatehgarh Sahib*
[2]*System Administrator, Computer Center, Mata Gujri College, Fatehgarh Sahib*

***Abstract- Energy benchmarking of campus buildings has become an important research area as universities seek to reduce operational costs, improve sustainability, and support carbon-reduction goals. This review examines data-analytics approaches used for benchmarking campus building energy performance, with attention to methods, metrics, datasets, and research challenges. The literature shows that benchmarking studies often rely on energy use intensity, weather-normalized indicators, regression analysis, clustering, forecasting, and machine learning to compare buildings and identify inefficiencies. Common data sources include smart meters, utility bills, occupancy information, weather data, and building management systems. The review also highlights major challenges such as data quality, lack of standardized benchmarks, limited occupancy-aware modeling, and few campus-specific studies in developing contexts. Based on the review, future research should focus on real-time analytics, explainable AI, integrated sustainability indicators, and benchmarking frameworks suitable for university campuses.***

***Keyword Energy Benchmarking, Campus Buildings,Data Analytics, Machine Learning, Energy Use Intensity (EUI)***

## I. INTRODUCTION

Globally, the building and construction sector represents a critical leverage point for climate change mitigation, given that it accounts for approximately 36% of global final energy consumption and roughly 39% of energy-related carbon dioxide emissions [1]. Within this broader urban context, higher education campuses function essentially as micro-cities.

They house a highly diverse mix of building types—including high-intensity laboratories, residential halls, administrative offices, and sports facilities—which results in an exceptionally complex energy profile. As universities face rising operational costs and escalating institutional pressures to meet net-zero carbon goals, administrators require deep, granular insight into campus energy distribution [2].

Energy benchmarking serves as the foundational mechanism for this process. By tracking a building's energy performance relative to its historical data (internal benchmarking) or relative to peer facilities (external benchmarking), campus managers can effectively isolate poorly performing structures and optimize efficiency investments [3].

Historically, benchmarking relied on static, monthly utility bills and simple aggregated metrics. However, the modern deployment of Advanced Metering Infrastructure (AMI) and Building Management Systems (BMS) has unlocked vast streams of high-frequency temporal data [4].

Consequently, contemporary campus energy management is shifting toward a data analytics perspective. Rather than relying purely on descriptive baselines, researchers now leverage multivariate regression, unsupervised clustering, and predictive machine learning algorithms to isolate variables such as weather fluctuations and complex structural attributes [5].

Despite these advancements, widespread implementation remains constrained by significant bottlenecks. The sheer structural and functional heterogeneity of university facilities complicates standard comparisons [6].

Furthermore, standard data streams are heavily vulnerable to data quality issues, and existing models rarely integrate real-time occupancy dynamics effectively [7].

Crucially, the current literature exhibits a severe geographic bias, with very few campus-specific benchmarking studies focused on the unique climate, financial, and institutional constraints found in developing contexts [3].

Objectives of this Review

To bridge these domains, this paper provides a comprehensive review of the state-of-the-art in data-driven energy benchmarking for higher education environments. Specifically, this study aims to:
1. Categorize the dominant data analytics methods and metrics used to benchmark campus facilities.
2. Characterize the primary data streams and integration frameworks powering these analytics pipelines.
3. Systematically evaluate the critical research gaps and structural vulnerabilities currently limiting the field.
4. Chart actionable pathways for future research, emphasizing real-time analytics, explainable AI (XAI), and regional equity.

## II. LITERATURE REVIEW

The existing body of literature on campus energy benchmarking sits at the intersection of building data science, statistical modeling, and institutional facilities management. This section synthesizes previous research across three main pillars: core benchmarking metrics, data-driven methodologies, and data sources.

### 2.1. Benchmarking Metrics and Indicators

The primary metric used across building energy literature is Energy Use Intensity (EUI), typically calculated as annual energy consumption normalized by a facility's total floor area [3]. While EUI serves as a useful foundational baseline, researchers widely acknowledge that raw EUI is insufficient for direct comparisons across a university campus.

For instance, a complex chemistry laboratory with specialized ventilation requirements inherently presents a vastly higher baseline energy demand than a humanities lecture hall or administrative block [5]. To resolve this limitation, studies frequently utilize statistical approaches—such as mean indexes, quartiles, and percentiles—to establish standard practical EUI ranges for distinct building classes [3].

Additionally, weather-normalization techniques using Heating Degree Days (HDD) and Cooling Degree Days (CDD) are widely applied to decouple ambient climatic conditions from a building's underlying operational efficiency [8].

To capture multidimensional dynamics beyond simple square footage normalization, advanced studies explore Data Envelopment Analysis (DEA) paired with Monte Carlo sampling to inject stochastic, multi-input, and multi-output constraints into the calculation of empirical efficiency frontiers [9].

### 2.2. Data Analytics & Machine Learning Methodologies

The literature highlights a clear evolution from traditional descriptive statistics to complex predictive and data-mining models [10]. The primary methodologies identified in contemporary studies fall into three distinct analytical categories:

- Statistical and Regression Modeling: Multiple Linear Regression (MLR) models and change-point formulations have historically formed the backbone of regional building benchmarking systems due to the ease of interpreting their coefficients [5]. Researchers utilize linear regression to establish basic energy baselines and quantify energy savings achieved post-retrofit [8]. However, linear formulations struggle to capture the highly complex, non-linear dependencies inherent to modern building operations.

- Unsupervised Learning (Clustering): Given the poor customizability of standard building classification codes for unique university facilities, researchers increasingly deploy clustering algorithms like K-Means, hierarchical clustering, or DBSCAN. These techniques allow analysts to automatically reclassify campus buildings based on actual energy usage intensities and behavioral load profiles, rather than relying strictly on nominal functional labels [5, 10].

- Supervised Machine Learning: To handle complex multi-variable interactions, recent studies deploy advanced non-linear architectures, such as random forests, support vector machines (SVM), and Artificial Neural Networks (ANN) [10]. For example, random forest methods are regularly leveraged to isolate specific building features that make the most significant contributions to total energy intensity [5]. More recently, Automated Machine Learning (AutoML) frameworks have emerged to systematically pick optimal model hyperparameters with minimal manual tuning, introducing a more scalable benchmarking execution layer [11].

2.3. Data Sources and Integration

Modern data-driven benchmarking requires the convergence of multiple, disparate data streams [4]. The literature identifies three core data layers required to feed analytics pipelines:

1. High-Frequency Metering Data (AMI): Smart meters provide the temporal data (e.g., 15-minute or hourly logs) required to assess sub-hourly load variations, flag sudden consumption anomalies, and monitor peak demand periods [4].
2. Building Management Systems (BMS): BMS layers yield localized operational data, including HVAC setpoints, indoor air temperatures, and fan speed modulations, which directly affect real-world equipment efficiencies [4, 7].
3. Contextual Data Streams: These include localized weather feeds (ambient temperature, humidity, solar irradiance) alongside institutional schedules (academic calendars, enrollment numbers, and classroom timetables) [3, 8].

## III. IDENTIFIED RESEARCH GAPS

Despite technical advancements in data engineering, the literature reveals four persistent gaps that impede the progress of campus-specific energy benchmarking:

1. The "Data Rich, Information Poor" Paradox: Although modern campuses gather immense volumes of data via smart sub-meters and BMS layers, much of this data suffers from severe quality issues [4]. Gaps due to uncalibrated sensors, telemetry failures, and a lack of unified semantic metadata standards (such as Project Haystack or Brick Schema) mean that researchers spend disproportionate effort cleaning data rather than extracting actionable operational insights.

2. Absence of Dynamic Occupancy Integration: The majority of existing benchmarking frameworks treat building occupancy as a static variable derived from fixed schedules or design capacity. Because student and faculty movement is highly fluid, treating occupancy as static introduces a massive "energy performance gap" between simulated predictions and actual measured loads [6]. The literature lacks robust frameworks that integrate real-time proxy metrics—such as Wi-Fi connection counts or localized indoor carbon dioxide sensors—directly into benchmarking models.

3. The Explainability Dilemma of Black-Box Models: While advanced machine learning architectures significantly reduce prediction errors compared to traditional linear models, they function as "black boxes" [10]. Facilities managers and university engineers are often hesitant to authorize capital-intensive retrofits or operational adjustments based on complex algorithmic outputs that cannot be transparently explained or visually traced [7]. Emerging literature points toward SHAP (SHapley Additive exPlanations) values to resolve this, but its application in campus systems is still in its infancy [11].

4. Geographic and Contextual Underrepresentation: The existing literature is heavily skewed toward institutions located in high-income regions across North America and Europe. There is a critical deficit of standardized energy benchmarking methodologies tailored specifically for the extreme climates, distinct institutional funding paradigms, and unique structural behaviors of university campuses within developing nations [3].

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1719689

## IV. PROPOSED FRAMEWORK FOR ADVANCED CAMPUS BENCHMARKING

To bridge the critical research gaps identified in Section 3, this section introduces a cohesive, multi-layered theoretical framework designed specifically for higher education environments. University campuses operate as socio-technical micro-cities where physical systems, human behaviors, and institutional schedules intersect.

Therefore, a robust benchmarking framework cannot rely on isolated data models; it requires a structured, multi-tier architecture that transforms raw data into transparent, actionable engineering decisions.

The proposed framework consists of four interdependent layers: the Physical and Data Acquisition Layer, the Semantic Data Governance Layer, the Analytical and Machine Learning Engine, and the Explainable Decision Support Layer.

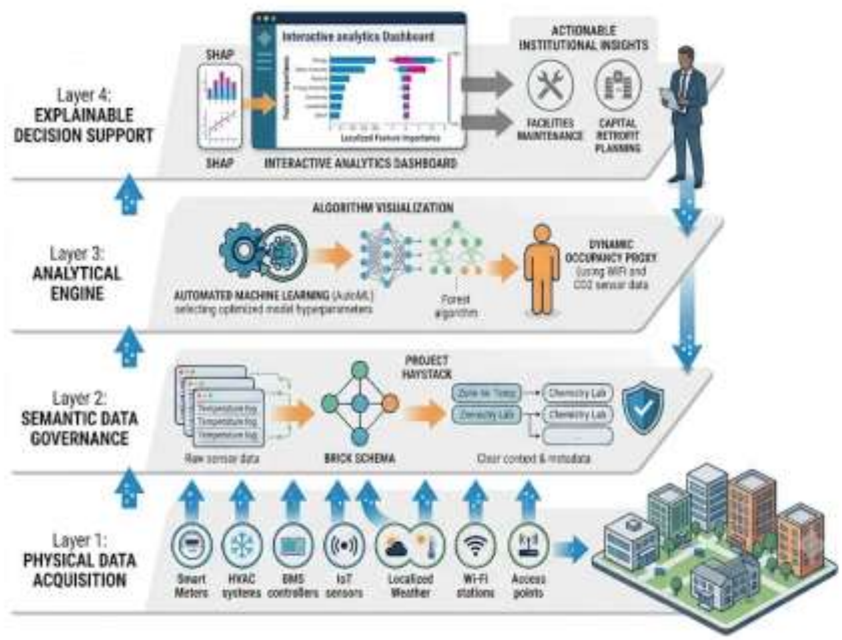


Figure1 Proposed Four-Layer Framework for Advanced Campus Energy Benchmarking

The architecture illustrates the end-to-end data pipeline: Layer 1 handles multi-source heterogeneous data acquisition (AMI, BMS, and non-intrusive Wi-Fi/$CO_2$ occupancy proxies); Layer 2 applies semantic metadata governance (Brick Schema/Project Haystack) to standardize sensor data; Layer 3 executes advanced non-linear predictive modeling via AutoML; and Layer 4 leverages SHAP interpretability to translate "black-box" models into actionable diagnostic insights for facilities management and capital retrofit planning.

### 4.1. Layer 1: Physical and Data Acquisition Layer

The foundational layer of the framework coordinates the ingestion of high-frequency temporal data and contextual streams. Unlike commercial real estate, a university's data infrastructure is decentralized and highly fragmented [4]. This layer partitions data streams into three major pipelines:

- High-Frequency Metering (AMI): Sub-hourly electricity, gas, and chilled water consumption logs to capture peak demand dynamics [4].
- BMS and Sensor Networks: Spatial variables including HVAC supply air temperatures, variable air volume (VAV) positions, and indoor $CO_2$ concentrations [7].
- Contextual Streams: Real-time localized weather metrics (ambient temperature, solar irradiance) combined with fluid institutional data (academic calendars and network Wi-Fi connection counts serving as dynamic occupancy proxies) [3, 8].

4.2. Layer 2: Semantic Data Governance Layer

To resolve the "Data Rich, Information Poor" paradox, the framework implements an abstraction layer utilizing open-source semantic metadata standards, specifically Brick Schema or Project Haystack [12].

Without semantic formatting, data algorithms spend excessive execution cycles on manual data wrangling due to naming discrepancies among different vendor BMS installations [4]. By mapping data points to a standardized ontology, a sensor label is programmatically contextualized.

This normalization automates data cleaning, flags telemetry dropouts, and enables benchmarking algorithms to scale seamlessly across an entire heterogeneous campus portfolio [13].

4.3. Layer 3: Analytical and Machine Learning Engine

The core computational layer processes the standardized data streams through advanced analytics engines. Rather than relying on static structural evaluations, this layer introduces multi-dimensional modeling:

- Dynamic Occupancy Normalization: By using localized Wi-Fi access point connection counts and metabolic $CO_2$ metrics as dynamic proxies, the framework replaces fixed occupancy schedules. This closes the "energy performance gap" by evaluating actual user-driven energy loads against real-time baseline expectations [2, 6].
- Advanced Non-Linear Modeling: Leveraging Automated Machine Learning (AutoML) routines allows for automated hyperparameter tuning and model selection (e.g., Random Forests, Support Vector Machines, Deep Neural Networks) to accurately model complex HVAC and thermodynamic behaviors across highly variable building types like high-intensity laboratories and administrative facilities [14].

4.4. Layer 4: Explainable Decision Support Layer

The ultimate objective of energy benchmarking is to drive energy conservation measures (ECMs) and support institutional capital retrofits. To prevent the user-adoption resistance typically caused by "black-box" machine learning pipelines, this framework integrates an interpretability layer powered by SHAP (SHapley Additive exPlanations) values [11].

By calculating the localized feature importance for every benchmarking evaluation, the framework translates complex model outputs into transparent operational diagnostics.

For example, instead of merely flagging a facility with a low efficiency score, the system provides campus engineers with a visual breakdown showing that the performance drop was driven 45% by simultaneous heating and cooling overrides in the BMS, 35% by unexpected occupancy during an academic holiday, and 20% by extreme ambient humidity.

This transparency bridges the gap between algorithmic predictions and physical facilities management [7, 10].

4.5. Framework Adaptability Matrix for Regional and Financial Contexts

Table 1: Framework Adaptability Matrix for Regional and Financial Contexts

| Implementation Tier | Data Infrastructure Availability | Core Analytical Methodology | Targeted Institutional Outcome |
|---|---|---|---|
| Tier 1: Low-Resource *(Developing Contexts)* | Monthly utility invoices, basic structural parameters, macro weather feeds [3]. | Weather-normalized EUI regression, statistical quartile profiling [5, 8]. | Identification of macro-level outliers; baseline institutional energy targeting. |
| Tier 2: Mid-Resource *(Transitioning Infrastructure)* | Smart meter AMI (hourly logs), academic timetables, static occupancy estimates [4]. | Unsupervised clustering (K-Means/DBSCAN) for diurnal load shape analysis [10]. | Peak demand shaving; optimization of operational schedules based on building types. |
| Tier 3: High-Resource *(Advanced Infrastructure)* | Full BMS integration, dense IoT sensor arrays, real-time Wi-Fi tracking proxies [7]. | Deep Learning / AutoML pipelines coupled with SHAP interpretability [11]. | Real-time automated anomaly detection; predictive maintenance; verified capital retrofits. |

The transition from descriptive, static energy benchmarking to advanced, data-driven analytics offers higher education institutions an unprecedented opportunity to operationalize sustainability. However, as the proposed framework demonstrates, achieving this shift requires overcoming a series of technical and systemic barriers.

A critical discussion point is the dependency on infrastructure equity. While Tier 3 implementations offer highly granular, real-time insights, they assume an institutional maturity in data engineering that is currently missing in many developing nations.

For resource-constrained universities, the focus must initially center on Tier 1 architectures—leveraging basic weather-normalization and statistical profiling—before committing capital to dense Internet of Things (IoT) or Advanced Metering Infrastructure (AMI) expansions [3, 5].

Furthermore, the integration of Explainable AI (XAI) addresses an ongoing sociological barrier within facilities management [11].

Engineering teams are traditionally risk-averse, relying on deterministic physics-based models. Introducing transparent algorithms through SHAP values can foster the institutional trust required to validate capital-intensive deep retrofits, proving that data science is an extension of engineering domain knowledge rather than a replacement for it [7, 10].

## V. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

This paper reviewed data analytics applications for energy benchmarking in campus buildings, highlighting the transition from static monthly Energy Use Intensity (EUI) calculations to high-frequency machine learning pipelines.

While advanced statistical models and predictive architectures have significantly enhanced anomaly detection and load profiling, widespread institutional adoption remains heavily constrained.

Key systemic bottlenecks include poor semantic data governance across disparate building management systems, a historical reliance on static occupancy assumptions, the "black-box" opacity of complex algorithms, and a deep geographic imbalance favoring high-income regions.

To overcome these vulnerabilities, future research must operationalize integrated, multi-tier frameworks that prioritize scalability and trust. Key technical priorities include adopting unified semantic metadata standards like Brick Schema to automate data integration [13], utilizing non-intrusive proxies like Wi-Fi connection logs for dynamic occupancy tracking [14], and embedding interpretability tools like SHAP to provide facilities managers with transparent diagnostics.

Expanding these adaptive methodologies to resource-constrained developing nations will ensure regional equity, enabling global university networks to leverage explainable AI in meeting institutional net-zero targets.

## REFERENCES

[1] H. M. Alshuwaikhat and I. Abubakar, "An integrated approach to achieving campus sustainability: assessment of the current campus environmental management practices," Journal of Cleaner Production, vol. 16, no. 16, pp. 1777-1785, 2008.

[2] J. Li, L. D. Iulo, and U. Poerschke, "A Review of the Energy Performance Gap between Predicted and Actual Use in Buildings," in Proceedings of the 18th IBPSA Conference on Building Simulation, Shanghai, China, 2023, pp. 1973-1980.

[3] M. A. Mohd Shukri, J. Jailani, and A. Hauashdh, "Benchmarking the Energy Efficiency of Higher Educational Buildings: A Case Study Approach," International Journal of Energy Economics and Policy, vol. 12, no. 2, pp. 491-496, 2022.

[4] Z. Afroz, H. B. Gunay, and W. O'Brien, "A review of data collection and analysis requirements for certified green buildings," Energy and Buildings, vol. 226, p. 110367, 2020.

[5] T. Wu, B. Wang, Zhang, D. Zhao, and H. Zhu, "Benchmarking Evaluation of Building Energy Consumption Based on Data Mining," Sustainability, vol. 15, no. 6, p. 5211, 2023.

[6] P. de Wilde, "The gap between predicted and measured energy performance of buildings: A framework for analysis," Automation in Construction, vol. 41, pp. 40-49, 2014.

[7] G. R. S. Santos, L. M. Honório, and A. L. M. Marcato, "Machine Learning for Energy Management in Buildings: A Systematic Review on Real-World Applications," Energies, vol. 19, no. 1, p. 219, 2026.

[8] S. Chung and W. Rhee, "A multi-period energy benchmarking model for university buildings using dynamic data," Energy and Buildings, vol. 84, pp. 411-421, 2014.

[9] Y. Hong, Y. Zhang, and J. M. S. Kim, "Objective Building Energy Performance Benchmarking Using Data Envelopment Analysis and Monte Carlo Sampling," Sustainability, vol. 9, no. 5, p. 780, 2017.

[10] S. A. Mansouri, T. S. Hasan, and M. K. Roy, "Systematic Review of Deep Learning and Machine Learning for Building Energy," Frontiers in Energy Research, vol. 10, p. 786027, 2022.

[11] X. Liang, Y. Chen, and C. Fan, "Benchmarking Automated Machine Learning for Building Energy Performance Prediction: A Comparative Study with SHAP-Based Interpretability," Buildings, vol. 16, no. 1, p. 185, 2026.

[12] J. Ploennigs, R. J. Miller, and M. Berges, "Semantic Web Technologies for Realizing the Digital Twin in Smart Buildings: A Review," Automation in Construction, vol. 142, p. 104492, 2022.

[13] S. Balaji, A. Bhattacharya, G. Fierro, J. Gao, J. Gluck, and Y. Agarwal, "Brick: Towards a Unified Metadata Schema for Buildings," Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments, pp. 41-50, 2016.

[14] A. Singh, H. B. Gunay, and K. Shen, "Extracting occupant presence and mobility profiles from university Wi-Fi connection logs for building energy simulation," Energy and Buildings, vol. 251, p. 111340, 2021.