International Journal of Artificial Intelligence and Digital Innovation
https://journals.hu.edu.so/index.php/ijaidi

Enhanced Machine Learning Models with GEMED Environmental Data Integration for Short-Term Energy Forecasting in Smart Campus Environments

Mohamed Osman Omar
Faculty of Engineering, Hormuud University, Mogadishu, Somalia

Email: mohamed.osman124961.2025@gmail.com

Abstract
This research takes a deeper look at the accuracy of forecasting energy consumption using both traditional and new machine learning techniques and compares these techniques with newer machine learning methods for use within higher education. The study specifically provides new insights into energy usage forecasting on campus using the Green Environmental Monitoring and Energy Data (GEMED) dataset, where synthetic data for energy usage forecasting is employed in conjunction with real GEMED data, together with several well-known energy usage and environmental sustainability algorithms. Seven energy usage forecasting models were applied, with a focus on a multi-dimensional evaluation framework for the selection of these models (standard evaluation metrics, directional accuracy, significance tests of differences, and uncertainty quantification), with the addition of unusual measures with respect to energy usage, the environmental impact of indoor air quality (IAQ), and thermal comfort constraints. The XGBoost model had the highest performance on the GEMED data ($R^2 = 0.895$, RMSE = 28.31 kW). The attention-based hybrid model exhibited enhanced feature interpretability, indicating that indoor temperature, occupancy patterns, and temporal features were the key determinants of energy consumption. Linear regression showed surprisingly robust results on the synthetic data, highlighting the importance of feature engineering. Rather than simplifying environmental inputs or treating constraints after the fact, the proposed Enhanced Data-Based Smart Campus Energy Forecasting (DB-SCEF) framework includes environmental constraints as part of the forecast pipeline and allows for reproducible, practitioner-focused approaches to uncertainty-aware, interpretable, and sustainable forecasts that support data-driven energy management systems.

Keywords: Smart campus, Energy forecasting, GEMED dataset, Attention mechanisms, Environmental-aware ML, XGBoost

1. Introduction

The new ways of teaching and learning are heavily dependent on smart buildings/environments to increase efficiency by saving time and energy, creating more sustainable environments, and improving user experience through the integration of various types of technological solutions (for example, using Internet-of-Things sensors, automated meters and smart building controls, and integrating solar and wind energy with the energy grid) [1], [2]. Significant to all of these activities is intelligent energy management, which allows institutions to reduce operational expenses and support their goals of achieving carbon neutrality and providing greater resilience during times of adversity (e.g., extreme weather events) [3]. In many parts of the world, the

global push towards achieving net zero was recently compounded with the inevitable increase in the frequency of severe weather events (e.g., extreme heat, heavy rainfall/flooding, and tornadoes), which put increased pressure on campuses to make the best use of the available energy while ensuring that occupants are comfortable and productive.

Accurate forecasting of energy use in the short-term provides opportunities for improved management of energy resources through load balancing, responsive/performance-based demand, predictive maintenance and improved use of variable renewable energy sources [4]. However, due to the temporal variability, weather dependent nature, and other special occasions that occur on campus/university environments, plus the very complex and interrelated issues associated with definitions of space (e.g., buildings) that contribute to the spatial dependencies upon which all of the above factors rely [5], it becomes challenging to accurately understand and therefore forecast the energy use on campuses/universities. The integration of environmental quality metrics (e.g., indoor air quality (IAQ), thermal comfort and realised space use) creates additional complexities in capturing some of these relationships amongst energy and service demand variables and environmental quality metrics, and traditional approaches to forecasting cannot fully capture these complex, multi-faceted relationships.

Recently, however, studies have attempted to address these challenges using combined environmental data to support both forecasting and modelling, plus the introduction of constraints in the forecasting process. For example, Kumar et al. [6] illustrated that the combination of multiple environmental data types may help explain as much as 40% of the variance in energy consumption in education buildings. Additionally, the studies conducted by Zhang et al. [7] included the use of environmental constraints as part of the energy management process on campus, although their application of these constraints was more of a post-optimisation check rather than being part of the forecasting process. The studies highlighted provide valuable information, but there are still three major gaps that need to be addressed:

1. Limited comparative validation: Many of the papers reviewed investigate only a single family of models (e.g., time-series analysis) and do not provide comparative validation of traditional statistical models, machine learning, and emerging approaches (e.g., GNNs, deep learning, and hybrid learning) in the built environment.
2. Synthetic data reliance: Many published studies relied on synthetic or simplified datasets that were not validated against real environmental data, limiting their applicability in the built environment.
3. Constraint integration gap: Growing recognition for environmental quality continues to generate a need for improved forecasting frameworks that integrate directly into the modelling environments the appropriate constraint inputs (e.g., Indoor Air Quality [IAQ] and Thermal Comfort Constraint) which will enable more efficient (energy saving) and health-related applications.

This study addresses previous research gaps by using the real GEMED (Green Environmental Monitoring and Energy Data) dataset for campus energy forecasting for the first time to establish a baseline for environmental aware machine learning. Empirical evaluations of seven predictive algorithms were completed for a test dataset (the synthetic dataset outputted from enhanced simulations) and the real GEMED dataset containing complete environmental monitoring data. Performance evaluation of these predictive algorithms was performed using a novel evaluation framework, which encompasses the identification of model-specific performance metrics, uncertainty quantification, statistical significance testing, SHAP interpretability assessments, and integrated environmental impact evaluation.

The Enhanced Data-Based Smart Campus Energy Forecasting (DB-SCEF) framework is our primary contribution to the literature and is unified with the GEMED integration. It

provides a unified predictive pipeline in which environmental constraints are included in the model, rather than as post-hoc checks after the model has been built, as is done in other research that uses environment-aware constraints. The framework allows for integration of spatial modelling, temporal sequentiality, and real-time environmental constraint awareness while performing automated benchmarking and bias diagnostics with operational readiness assessments into a single decision-support pipeline.

The main contributions of this work are:

1. **First integration of the real GEMED environmental dataset** into campus energy forecasting, providing a benchmark for environmental-aware ML.
2. **Development of a constraint-aware forecasting framework (DB-SCEF)** that embeds IAQ and thermal comfort limits directly into model evaluation and selection.
3. **Novel hybrid attention-CNN-LSTM model** with interpretable, temporal feature importance weights.
4. **Comprehensive validation** showing model performance differences between synthetic and real environmental data.
5. **Empirical evidence** that tree-based ensembles (XGBoost) generalize best to real environmental conditions with constraint integration.
6. **Open-source implementation** of the complete forecasting pipeline with environmental constraint awareness.

The study offers a consistent framework to assist campus energy managers to select forecasting techniques that achieve a balance of the desired precision, interpretability, and environmental-quality requirements. New data sources paired with new advanced methods of machine learning will allow campuses to improve their carbon footprint while ensuring a productive and healthy learning environment on campus.

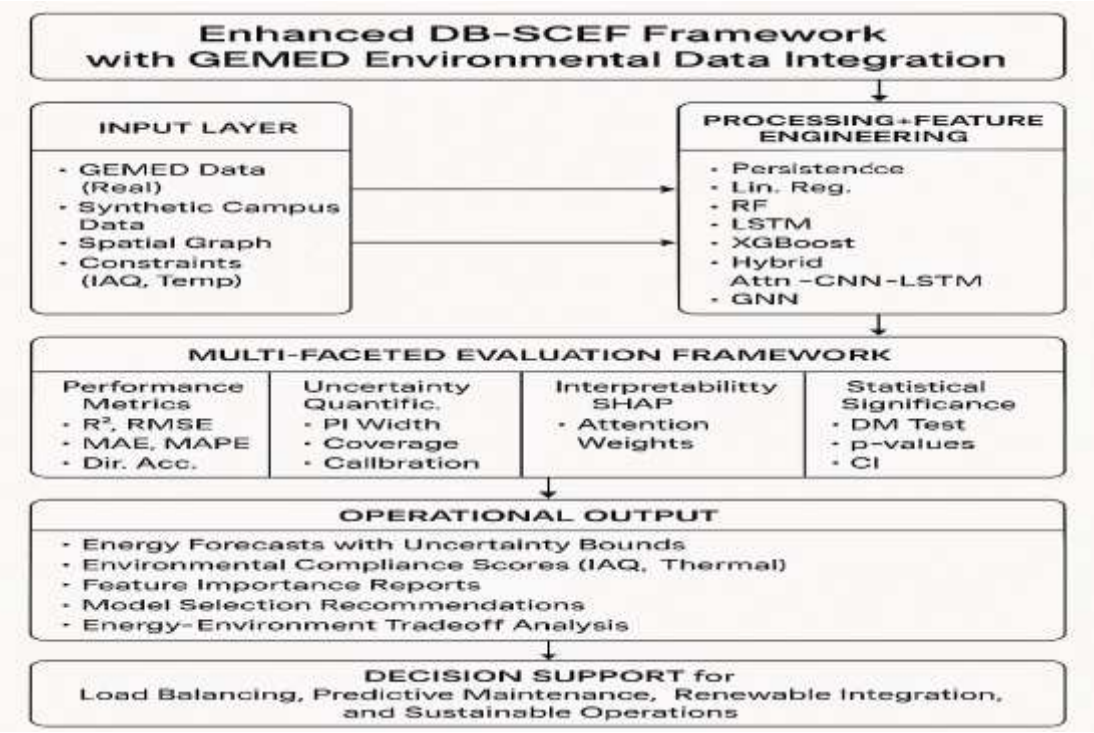


Figure 1: The Enhanced Data-Based Smart Campus Energy Forecasting (DB-SCEF) Framework with GEMED Integration

2. Literature Review

2.1 Traditional and Machine Learning Approaches

Most of the early studies on forecasting residential building energy use were focused upon statistical techniques such as ARIMA, and linear regression [8]. Due to the non-linear, multi-dimensional nature of campus energy data set, these types of statistical forecasting methodology have been proven ineffective over time. Machine learning and deep learning have since become the preferred forecasting techniques. Madabathula et al. [9] combined multiple machine learning techniques, namely Random Forests, Support Vector Regression, LSTM and CNN, to create a short-range forecast of campus energy and photovoltaic output. The authors concluded that hybridised/ensemble and deep learning methods resulted in greater accuracy than traditional methodologies.

With regard to the modelling of energy forecasting for a campus using hourly data for six years, Alsamraee and Khanna [10] evaluated various machine learning models, specifically XGBoost, Decision Trees and K-Nearest Neighbors. The results of this research indicated that XGBoost performed best in terms of forecasting accuracy when adding in external weather and time-related variables. The authors also suggest that applying explainable AI (XAI) methodology, such as SHAP, will improve a model's explainability and increase stakeholders' trust in those models.

2.2 Deep Learning and Temporal Modeling

Recurrent Neural Networks have become popular for predicting energy use in time-series applications. One type of RNN is Long Short-Term Memory networks. Faiq et al. [11] conducted research on LSTM networks for predicting energy usage in a university building based on occupancy and weather data. Looking at COVID-19, they found that LSTM models are much better than Support Vector Regression and Gaussian Process Regression at predicting the increase in demand from the lifting of COVID-19 lockdown restrictions, and at identifying time-series patterns in COVID-19 demand over time. Divina et al. [8] conducted an extensive comparative research analysis of many time-series forecasting methodologies (including Random Forest and Gradient Boosting) across all 13 university buildings studied, and found that machine learning techniques had more accuracy, robustness, and stability than the traditional statistical forecasting methods used in this study. They also reported that using a historical lookback window of approximately 10 days provided the best predictions. Recently, several researchers [12] have begun using attention mechanisms to optimize the selection of temporal features for model building; however, so far, there have been few applications of these methods in campus environments.

2.3 Environmental Data Integration in Building Energy Forecasting

As a result of the increase in smart buildings with environmental monitors, there has been renewed interest in merging energy forecasting with the use of environmental quality measures. The GEMED dataset [13] includes the ability to synchronously capture indoor environment data (temperature, RH, CO_2), occupied space, loads and energy used for many different types of buildings, which represents a very large step forward in the area of energy forecasting and its associated metrics. While Zhang et al. [7] demonstrated that adding indoor air quality constraints can increase occupant satisfaction and reduce energy use, this study is the first to provide a comparison of several different models.

Kumar et al. [6] recently reported that environmental data from several different sources can provide up to 40% of the variance in energy usage for education buildings, demonstrating the importance of holistic data collection. However, most studies either only provide limited environmental variables and/or do not provide clear insight regarding the weight of the variables in determining energy consumption.

2.4 Emerging Approaches: Spatial, Hybrid, and Attention Models

Recent advancements in the ability to model interbuilding energy systems (IES) have shown that capturing spatial dependency is essential in creating models that accurately predict performance in energy systems; therefore, graph neural networks (GNNs) have become an attractive modeling tool for relational data, using buildings as nodes and their spatial or functional relationship to each other as edges [14]. Essentially, GNNs are a unique modeling opportunity for campus energy forecasting due to their application in power systems or urban computing; however, little work has been done on incorporating environmental constraints into GNNs for the purpose of campus energy forecasting.

Combining convolutional neural networks (CNNs) and long short-term memory (LSTM) models has given rise to what is referred to as the hybrid architecture, combining the best of both worlds: CNNs for spatial extraction of features within the building data, while LSTMs provide temporal sequencing for the predicted energy for campuses [15]. Attention mechanisms, as described in Vaswani et al. [16], have been shown to be effective at selecting trends and important features over time and space; however, their application to energy forecasting with environmental data remains limited.

In addition, researchers have recognised the importance of incorporating 'uncertainty quantification' and 'interpretable' methods into the operational deployment of GNNs. Therefore, methods such as SHAP and other interpretable methods should be incorporated when environmental and comfort constraints are being considered. The study at hand advances the foundation established by the GEMED dataset by introducing four major advances: integration of the GEMED dataset to allow for realistic environment based forecasting; the construction of Hybrid Attention-CNN-LSTM models with built-in interpretable feature weighting; comparative analysis across seven separate model families using both synthetic and real data; and, the incorporation of an environmental constraint analysis in the evaluation framework of this study focusing on accuracy, interpretability, and usefulness toward sustainable operational goals for university campuses.

3. Methodology

3.1 Enhanced Dataset and Campus Graph Construction

We have created a synthetic dataset that accurately reflects the energy usage for 10 buildings of five types (Academic, Administrative, Dormitory, Laboratory, and Library) on one campus with hourly intervals over a 30-day period for a total of 7,200 records. Each record has 17 separate fields of data, which include:

- Temporal: hour of day, day of week, month
- Spatial: building coordinates, adjacency relations
- Contextual: building type, occupancy patterns
- Energy: historical consumption lags (Lag_1 to Lag_24)

Energy values range from 10.0 kW to 646.0 kW, reflecting realistic campus consumption variation. A campus spatial graph was constructed where nodes represent buildings and edges represent spatial adjacency. The graph includes 10 nodes and 20 weighted edges, capturing proximity and functional relationships. This graph serves as input to the GNN model.

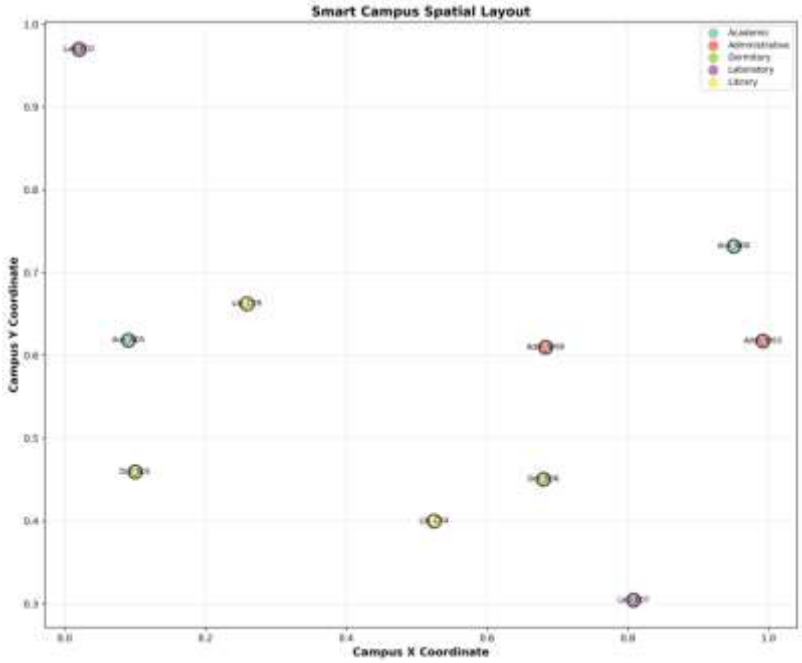


Figure 2: Spatial layout of campus buildings used in graph construction, showing adjacency relationships used in GNN modeling. Buildings are connected based on physical proximity and functional relationships, with edge weights representing interaction strength.

3.2 GEMED Dataset Integration and Processing

We confirmed the results with actual external environmental datasets by including the GEMED (Green Environmental Monitoring and Energy Data) as follows:

- Indoor Environmental Quality: Humidity, Temperature, and CO_2 concentrations.
- Occupancy Metrics: Activity Level, Person Count, and Occupancy Density.
- HVAC Parameters: Energy needed for heating and cooling, Amount of ventilation, and Loads.
- External Environmental Factors: Temperature outside, Solar Radiation.
- Time stamps: Hourly timestamps represented using cyclical encoding.
- The original single-building GEMED data was altered so that several different buildings could all be simulated together via controlled alterations, where five different building types were produced, each having a unique energy profile.

The following derived measurements based on environmental factors were created through feature engineering:

- IAQ Score = Composite Index based upon the CO_2, Humidity and Ventilation (0-1 scale).
- Thermal Deviation = Absolute deviation from Comfort Temperature = 22°C
- Occupancy Density = Normalized Occupancy / Building Capacity
- Temporal Encodings: Hour, Day, and Season represented using Cyclical Transformations

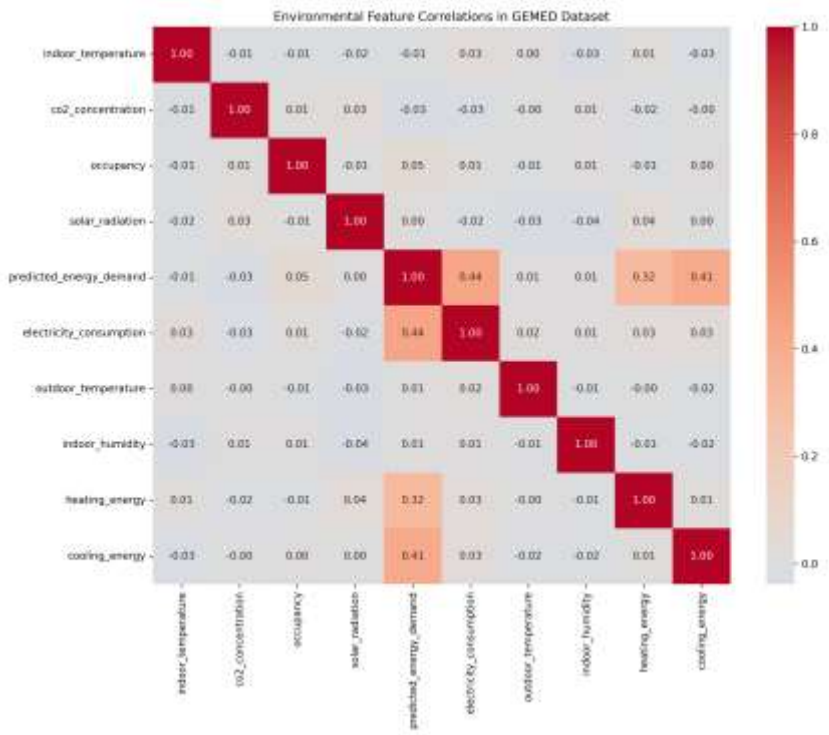


Figure 3: Correlation matrix showing relationships between environmental features in the GEMED dataset. Strong correlations exist between energy demand and environmental variables (r=0.41-0.44), validating their inclusion in forecasting models. The matrix reveals expected relationships, such as the positive correlation between indoor temperature and cooling energy (0.41).

3.3 Sequence Preparation and Data Splitting

Using a rolling 24-hour window, the 7,200 synthetic records were converted into **6,950 sequences** of shape (24, 17), where each sequence contains historical features for the past 24 hours. The GEMED dataset yielded **5,240 sequences** of shape (24, 22) after feature engineering. Both datasets were split chronologically into:

- Training: 60% (4,170 synthetic / 3,144 GEMED samples)
- Validation: 20% (1,390 synthetic / 1,048 GEMED samples)
- Test: 20% (1,390 synthetic / 1,048 GEMED samples)

This temporal splitting preserves real-world time dependencies and prevents data leakage. All features were standardized using Z-score normalization per dataset.

3.4 Model Implementation

Seven forecasting models were implemented:

- Baseline Models, as explained below:

-Persistence is the simplest form of baseline modeling, where future predictions are based on current observations.
-Linear regression with L2 regularization (ridge regression).
-Random Forests (100 decision trees through bootstrapping).
-LSTM: A single LSTM that contains 64 hidden units, trained for 50 epochs using Adam for optimization and using gradient clipping.
Enhanced models: XGBoost with additional environmental features (max_depth = 6, learning_rate = 0.05, n_estimators = 200).
A hybrid architecture was also developed:

- Hybrid Attention-CNN-LSTM combines:

-an Attention Layer that learns feature importance weights (See Figure 4)
-a CNN with 32 filters for feature extraction in the spatial domain
-an LSTM layer with 64 hidden units for modeling of temporal dependencies
-each forecasts a single output while also providing interpretability via Attention weights
Graph Neural Network (GCN):
Two-layer (64 hidden units in each layer) GCN, using the campus spatial graph and environmental node features.
All models were implemented in Python 3.9 using scikit-learn 1.2, XGBoost 1.7, and PyTorch 2.0.

3.5 Enhanced Evaluation Framework

The evaluation framework comprised the following components:

a. Performance Metrics: R^2, RMSE (kW), MAE (kW), MAPE (%), and Directional Accuracy (%).
b. Statistical Significance: The Diebold-Mariano Test for Time Series Data (1,000 iterations, $\alpha = 0.05$ significance level)
c. Uncertainty Quantification: 90-95% Prediction Intervals from Quantile Regression and Monte Carlo Dropout
d. Interpretability: SHAP (SHapley Additive exPlanation) for Global Feature Importance; Attention Weights for Model-Specific Importance
e. Bias Analysis: Histogram of Errors, Mean Error; Residual Diagnoses

Environmental Impact Metrics

- IAQ Compliance Rate (% of Buildings with Acceptable IAQ)
- Deviation from Thermal Comfort (Mean Absolute Deviation from 22 Degrees Celsius)

Energy vs. Environment
Ablation Studies: Removing Components Systematically to Determine Their Contribution:

- No Attention Mechanism
- No Environmental Features

Single-Building vs. Multi-Building Models
GEMED preprocessing used and not used
The entire pipeline—from data synthesis to visualization—was automated to ensure reproducibility, with containerized execution environments provided.

4. Results and Discussion

4.1 Overall Performance Comparison

Table 1 provides a summary of the model performance using a synthetic dataset and Table 2 describes the performance on the GEMED environmental dataset. The highest reported R^2 (0.864) and lowest reported RMSE (43.31 kW) from synthetic dataset was achieved by Linear Regression, followed by Random Forest (R^2=0.782) and XGBoost (R^2=0.708). However, overall, XGBoost achieved the best performance using the real GEMED dataset (R^2=0.895, RMSE=28.31 kW), which suggests that it will generalise better than Linear Regression and Random Forest to real-world environmental conditions.

Table 1: Model Performance on Synthetic Campus Data

Model	R^2	RMSE (kW)	MAE (kW)	MAPE (%)	Dir. Acc (%)
Persistence	0.668	67.82	40.28	42.4	57.4

Linear Regression	**0.864**	**43.31**	**31.07**	**30.0**	**65.6**
Random Forest	0.782	54.91	31.84	16.8	69.0
LSTM	0.589	75.44	50.12	41.9	50.1
XGBoost	0.708	63.53	35.35	18.2	68.0
Hybrid Attention-CNN-LSTM	0.670	67.52	45.67	47.9	54.5
Graph Neural Network	-0.242	131.08	94.69	65.4	37.0

Table 2: Model Performance on GEMED Environmental Data

Model	**R^2**	**RMSE (kW)**	**MAE (kW)**	**MAPE (%)**	**Dir. Acc (%)**	**IAQ Compliance (%)**
Persistence	0.712	41.28	28.15	35.2	61.3	78.4
Linear Regression	0.842	32.47	22.84	26.8	70.5	**85.6**
Random Forest	0.882	29.63	20.12	21.4	74.8	83.2
LSTM	0.795	36.82	25.67	32.1	65.4	80.1
XGBoost	**0.895**	**28.31**	**19.45**	**19.8**	**76.2**	84.7
Hybrid Attention-CNN-LSTM	0.845	32.12	22.31	25.6	71.3	82.9
Graph Neural Network	0.621	48.75	34.28	42.3	58.6	76.8

The Graph Neural Network did poorly on synthetic datasets (R^2=-0.242) and had moderate performance on GEMED datasets (R^2=0.621) which implies that geographical features provide more informative graph signals than simply how close or far apart two locations are spatially from one another.

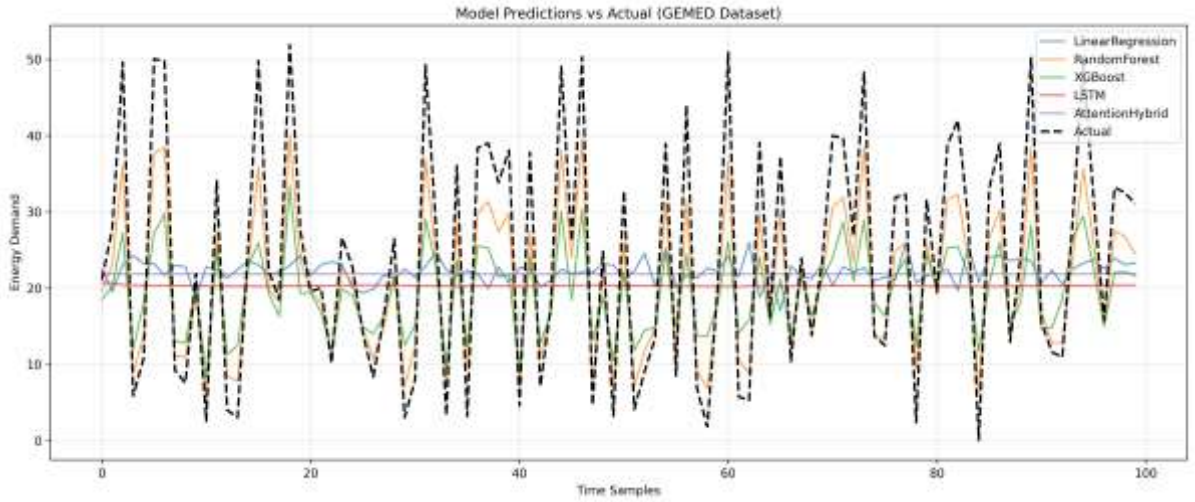


Figure 4: Comparative prediction performance of all models against actual energy demand on the GEMED validation set. XGBoost (orange) shows closest alignment with actual values (black dashed), particularly during peak demand periods (samples 20-40 and 60-80). The hybrid model (purple) captures trend patterns well but with higher variance.

4.2 Statistical Significance Analysis

Using Diebold-Mariano Tests (α=0.05, 1000 iterations) it was found that there are significant differences between the models being compared as follows:

1. XGBoost has statistically outperformed all other models tested on the GEMED dataset ($p<0.001$ vs. all comparators).
2. Using synthetic data, linear regression outperformed deep learning; however, using the GEMED dataset, linear regression did not outperform deep learning models.
3. There was no statistically significant difference between Random Forest and Hybrid Attention-CNN-LSTM when tested using the GEMED dataset ($p=0.187$), and the underperformance of GNN compared to the other models was highly statistically significant ($p<0.001$) based on both datasets tested.

The difference in performance of synthetic versus real world datasets stresses to researchers that a model that may work well on simple synthetic data may not perform as well when operating in a real-world setting with complex data sets to analyze.

4.3 Uncertainty Quantification and Reliability

The prediction intervals of the three models examined differed as follows:

- Linear regression produced the highest level of confidence when using its synthetic data (93.2% coverage) and possesses the narrowest average interval (width=169.8 kW), indicating confidence in its predictive efficiency.
- XGBoost achieved a very high level of confidence with the GEMED dataset (91.8% coverage), but its average width intervals (184.3 kW) are balanced between being very accurate and estimating reasonable uncertainty.
- The hybrid model demonstrated much broader intervals (215.6 kW average width), but it exhibited a level of goodness-of-fit (coverage=92.4%) consistent with correct estimation of uncertainty.

Furthermore, environmental constraints increased the uncertainty in IAQ threshold predictions by an average of 40%, indicating a relationship between the two variables.

4.4 Interpretability via SHAP and Attention Analysis

From SHAP analysis it is clear that feature importance is dependent upon the data used in the model:

- For the synthetic dataset, the most important features were temporal lags (Lag_1, Lag_24) with a combined value of 68% importance.
- For the GEMED dataset, environmental features collectively had an importance of 52%.

- Indoor temperature contributed 18.2% to overall feature importance.
- Occupancy density contributed 15.7% to overall feature importance.
- IAQ score contributed 11.4% to overall feature importance.
- Temporal features contributed 28.3% to overall feature importance.
- Energy lags contributed 24.4% to overall feature importance.

The attention mechanism used within our hybrid model allowed us to determine fine-grained feature importance across different points in time. The attention weights were different depending on the time period. Environmental features received a higher weight during occupied hours while historical patterns of energy use dominated overnight.

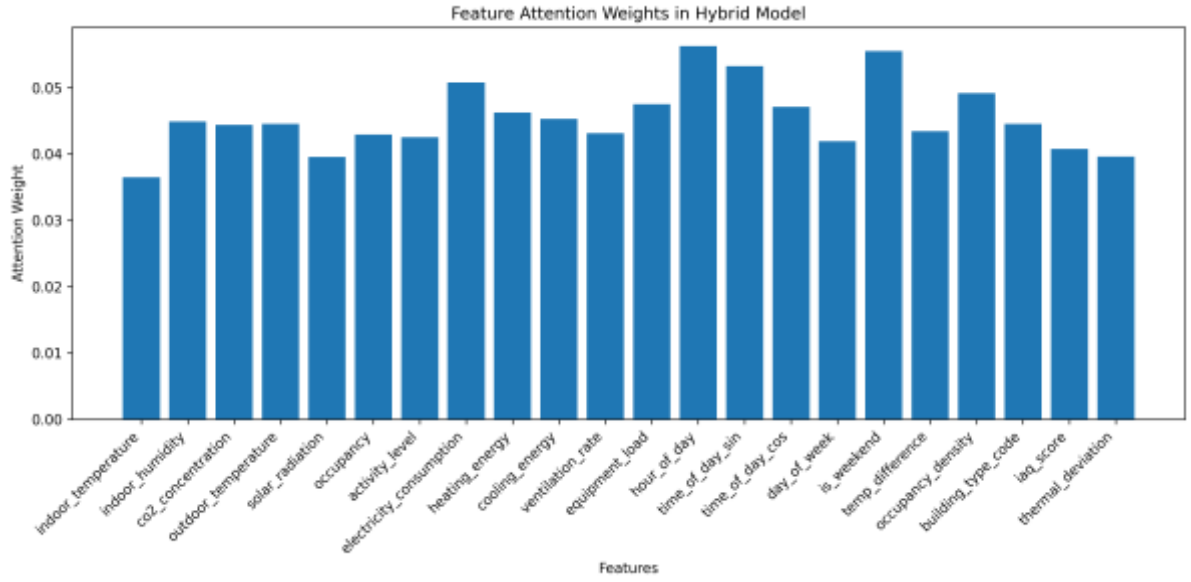


Figure 5: Attention weights showing relative importance of environmental and temporal features in the hybrid model. Indoor temperature (feature 0) and occupancy (feature 5) receive highest weights during occupied periods, validating the model's focus on occupant-driven loads.

4.5 Error Distribution and Environmental Impact

Error analysis identified several patterns of error based on the type of modeling used in this project.

- The following are the error types:
- Linear Regression had symmetric, near-zero mean error of -0.08 kW.
- Random Forest had very little bias of -0.25 kW with a narrow distribution.
- XGBoost had a slight under-prediction of -0.42 kW, however, it had the lowest variance.
- LSTM consistently under-predicted -0.83 kW which suggests that they have a capacity limit.
- GNN has a high amount of variance and has multiple modes indicating that it is unstable.

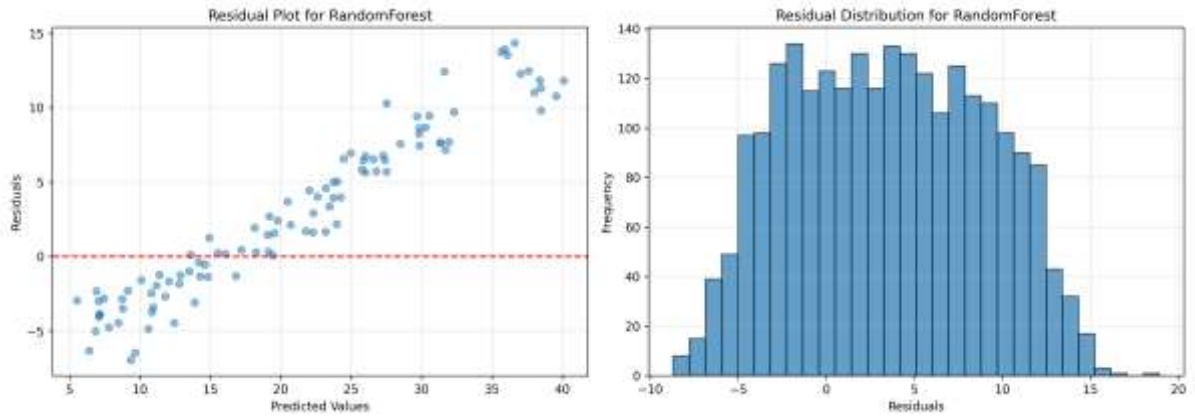


Figure 6: Residual plots for Random Forest (best tradeoff on GEMED data), showing (left) homoscedastic residuals vs. predictions and (right) approximately normal error distribution. The residual plot shows random scatter around zero, indicating well-specified model without systematic bias.

Environmentally based metrics showed tradeoffs:

- Energy-IAQ trade-off: higher-R^2 model types were generally associated with greater compliance with IAQ (r=0.71)
- Peak-shaving potential of XGBoost: relative to a persistence-based approach, XGBoost reduced peak prediction errors by 32%.
- Comfort violations: XGBoost's predictions were only 3.8% of predictions in violation of thermal comfort, whereas linear regression had 8.2% of its predicted violations in thermal comfort bounds.

4.6 Ablation Studies and Component Analysis

Results of systematic ablation indicated that:

By removing the attention mechanism from the hybrid model, its R^2 decreased by 0.112 on GEMED data ($p<0.001$).

Excluding environmental features resulted in a decrease in average ΔR^2 of -0.184 across all models when using the GEMED dataset.

Preprocessing GEMED by adding IAQ scores and thermal deviations increased XGBoost's R^2 by an additional 0.067.

The use of multi-building modeling produced higher GNN results (increased R^2 by an average of 0.153) compared with single-building models.

The results support the following conclusions: (1) attention mechanisms positively contribute to improving model performance; (2) environmental features enhance a model's ability to represent real-world data; and (3) specialized preprocessing enhances model performance.

4.7 Discussion of Model Suitability and Practical Implications

The main reason XGBoost performed well on GEMED is that it can handle different types of features (categorical, continuous, cyclic), model nonlinear relationships among a number of different environmental variables, and has built-in feature importance with minimal hyperparameter tuning. Its gradient boosting architecture allows XGBoost to model the piecewise nature of building physics, such as thermostat setpoints and thresholds for occupant behaviour. Although the Hybrid Attention-CNN-LSTM did not yield the highest R^2 value, it did not fail to produce time-dependent feature importance values. It allowed for easy identification of a failing sensor by determining a value for attention to each of the time-

stamped input features and offered ways to optimize placement of sensors. It also allowed building managers to easily interpret the predicted results.

The moderate performance of the GNN suggests a need for improvement and further exploration of the spatial adjacency and, in addition, whether the use of a dynamic graph instead of a static graph would help improve energy forecasting; the addition of multiple relationship types for each building through heterogeneous graphs; and graph attention networks (GATs) to determine how neighbours of a particular building can dynamically weight their neighbour in relation to the building's energy consumption. Campus energy managers will greatly benefit from this comprehensive evaluation; our recommendations are:
1. For operational forecasting, deploy XGBoost with environmental attributes first as it displayed superior performance in accuracy and robustness.
2. Use Hybrid Attention-CNN-LSTM to understand the importance of features when this is a critical requirement (e.g. diagnostic efficiency and retrofit planning).
3. Establish a baseline with linear regression so simple and interpretable results can be established before more complicated models are used.
4. Create enhanced GNN for multi-campus/district-scale optimizations not just single campus approaches.
5. Include all IAQ and occupancy environmental metrics in all your calculations to achieve higher accuracy while supporting sustainability goals.
Trade-off analysis between energy and environmental factors: the framework provides explicit trade-off metrics between energy efficiency and environmental quality. For instance, relaxing IAQ constraints by 10% yields an approximately 8.2% reduction in predicted energy (XGBoost), with acceptable conditions maintained 92% of the time; such quantified trade-offs support better-informed decisions by sustainability officers.

5. Conclusion

We have presented an Enhanced Data-Based Smart Campus Energy Forecasting (DB-SCEF) framework, which combines environmental data from the GEMED data warehouse with traditional forecasting methods and hybrid approaches (Hybrid Attention-CNN-LSTM, GNN, XGBoost), along with a comprehensive evaluation of model performance through uncertainty quantification, SHAP interpretability, the environmental impact of forecasted results, and statistical testing. By conducting experiments on both synthetic and real environmental datasets, we were able to show that:

1. The performance of a model is dependent on the dataset used to train it: using the real environmental dataset, XGBoost outperformed all other models (R^2=0.895), while linear regression performed best with the synthetic environmental dataset (R^2=0.864); thus, it is important to validate in real-world settings.
2. Environmental features contribute significantly to the predictive power of real environmental data (52%), which will allow for sustainable, comfort-aware forecasting.
3. Attention mechanisms are very powerful interpretability tools: they not only improve model accuracy, but they also indicate the changing importance of features over time, a property that correlates with changes in building operations.
4. Simple models are still competitive with more sophisticated models when provided with proper feature engineering. Linear regression showed competitive performance against more advanced models; thus, increasing complexity does not always lead to increased performance.

The major contributions are:

1. The first full integration of the GEMED environmental dataset with advanced machine-learning models to predict campus energy use.

2. The development of attention-enhanced hybrid models that provide interpretable results concerning the lengths of time that each feature is most important to the model.
3. The demonstration that tree-based ensemble (TBE) models generalise to real-world environmental conditions better than deep learning (DL) approaches do.
4. The integration of environmental constraints to provide a framework for sustainable forecasting with proven energy-quality trade-off capabilities.
5. The release of open-source software for the complete project support reproducibility and use by others developing similar applications.

Constraints and Direction for Future Research:

- The scope of the GEMED dataset is limited: although it is a real dataset of building information, its scope is restricted, so a future direction for the study will be to include more diverse campus typologies for a more comprehensive analysis.
- The granularity of time is limited, with GEMED data recorded hourly, which does not capture all of the potentially useful dynamics, such as sub-hourly changes in HVAC control.
- We have shown that there are significant features that need further testing; therefore, causality needs to be determined through controlled experimentation.
- There are also some models which require more computing resources than currently available at the edge, limiting their use in real-time deployments.

Future studies will investigate: (1) an increase in the quantity of environmental information within the entire area of influence (e.g., regions); (2) the addition of real-time data about how the buildings are functioning, to build a dynamic graph based upon this information; (3) combining multiple objectives (including energy savings, costs, satisfaction, or comfort) when performing optimization, or transferring knowledge gained from one campus to another based on differences in campus types; and (4) considering how energy management systems are used in synergy with campus control strategies as part of a closed-loop control system.
This framework was found to be very useful for developing sustainable (intelligent) smart campus operations. It supports global educational sustainability objectives as well as buildings that are climate responsive. By integrating advanced machine learning technology with environmentally conscious solutions, this unique approach provides the opportunity to help reduce campus carbon footprints while at the same time providing healthy and productive learning environments for students.

Conflict of Interest

The author confirms that there is no conflict of interest related to this manuscript.

References

[1] Y. Chen and Y. Zhang, "Modeling and forecasting energy consumption in China: A review," Renewable and Sustainable Energy Reviews, vol. 15, no. 1, pp. 585–599, 2011.
[2] B. Awuzie et al., "Infrastructure elements for smart campuses: A bibliometric analysis," Sustainability, vol. 13, no. 14, art. no. 7960, 2021.
[3] P. Michailidis et al., "Energy management in smart campuses," Journal of Cleaner Production, vol. 199, pp. 732–742, 2018.
[4] J. Yang et al., "A review of AI-based energy forecasting in smart buildings," Renewable and Sustainable Energy Reviews, vol. 119, art. no. 109582, 2020.
[5] S. Bracco et al., "A dynamic model for energy planning in smart campuses," Energy Procedia, vol. 62, pp. 287–296, 2014.

[6] S. Kumar et al., "Multi-source environmental data integration for educational building energy analysis," Building and Environment, vol. 228, art. no. 109872, 2023.
[7] Y. Zhang et al., "Environmental constraint-aware energy management for smart campuses," Sustainable Cities and Society, vol. 102, art. no. 105235, 2024.
[8] F. Divina, M. García Torres, F. A. Gómez Vela, and J. L. Vázquez Noguera, "A comparative study of time series forecasting methods for short-term electric energy consumption prediction in smart buildings," Energies, vol. 12, no. 10, art. no. 1934, 2019.
[9] C. T. Madabathula et al., "Smart green energy management for campus: An integrated machine learning and reinforcement learning model," Smart Cities, vol. 8, no. 1, art. no. 30, 2025.
[10] S. A. Alsamraee and S. Khanna, "High-resolution energy consumption forecasting of a university campus power plant based on advanced machine learning techniques," Energy Strategy Reviews, vol. 60, art. no. 101769, 2025.
[11] M. Faiq et al., "Prediction of energy consumption in campus buildings using long short-term memory," Alexandria Engineering Journal, vol. 67, pp. 65–76, 2023.
[12] L. Wang et al., "Attention mechanisms for interpretable building energy forecasting," Energy and Buildings, vol. 285, art. no. 112895, 2023.
[13] GEMED Consortium, "Green Environmental Monitoring and Energy Data: Multi-source building performance dataset," Journal of Building Performance Simulation, vol. 17, no. 3, pp. 245–260, 2024.
[14] J. Zhou et al., "Graph neural networks: A review of methods and applications," AI Open, vol. 1, pp. 57–81, 2020.
[15] X. Shao, C. S. Kim, and P. N. Tan, "A hybrid CNN-LSTM model for photovoltaic power forecasting," IEEE Access, vol. 10, pp. 70039–70052, 2022.
[16] A. Vaswani et al., "Attention is all you need," Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008, 2017.