

Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study

John Adejoh ^{1,✉}, Chibuzor Ezeamasiobi ¹, Ugonna Obi-Emeruwa ²

1. Department of Software Engineering, African University of Science and Technology, Abuja, NGA

2. Department of Public Policy and Public Administration, African University of Science and Technology, Abuja, NGA

Received: August 10, 2026 | Review began: August 19, 2026 | Review ended: September 26, 2026 | Published: October 28, 2026

© Copyright 2026

This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

The escalation of global building energy consumption, driven substantially by heating, ventilation, and air conditioning (HVAC) systems, necessitates robust, privacy-preserving analytics for energy prediction and fault detection. However, building meter data is fragmented across owners and vendors, preventing centralization of raw data for model training. While federated learning offers a solution, systematic comparisons under real-world cross-climate heterogeneity remain limited. This paper presents a proof-of-concept federated learning framework evaluated on a representative subset of five diverse buildings from the Building Data Genome Project 2 dataset. We integrate Federated Averaging (FedAvg), FedProx, climate-zone Clustered Federated Learning, and Personalized Federated Learning for next-hour HVAC load prediction, alongside a federated autoencoder for anomaly detection. Comparative analysis reveals that flat FedAvg is consistently the weakest prediction strategy [coefficient of variation of the root mean square error (CV (RMSE) = 9.99%)], while Clustered and Personalized Federated Learning (FL) recover most of the accuracy gap, with Personalized FL achieving the best overall performance (CV (RMSE) = 7.98%, $R^2 = 0.861$). For fault detection using synthetically injected fault signatures (sensor drift, stuck-meter, and spike patterns), the federated autoencoder achieves a superior F1-score (0.794 vs. 0.748) compared to a local-only detector. Notably, both detectors exhibit near-identical receiver operating characteristic-area under the curve (~ 0.972), indicating that threshold calibration rather than representational capacity drives the observed operating differences. This preliminary study underscores the efficacy of federated learning as a privacy-preserving tool, establishing a baseline for future large-scale validation across full building portfolios.

Categories: Data Analysis, Machine Learning (ML), Deep Learning

Keywords: federated learning, building energy prediction, fault detection, data heterogeneity, climate zones, building energy management, deep learning

Introduction

The building sector has become one of the largest and fastest-growing consumers of energy worldwide. Buildings account for roughly 30% of global final energy consumption and over half of global electricity consumption [1]. This demand is accelerating rather than plateauing: global electricity consumption in buildings rose by more than 600 TWh in 2024 alone, a 5% increase that accounted for nearly 60% of total growth in electricity consumption that year [2], driven largely by rising air-conditioning demand amid more frequent heatwaves and growing cooling needs in countries such as India and China [3]. Within this footprint, heating, ventilation, and air conditioning (HVAC) systems are consistently the single largest end use, typically accounting for around a third of a commercial building's total energy consumption, depending on climate and building type [4], with Department of Energy reporting that HVAC systems account for approximately 30% of primary energy use in commercial buildings, and that heating and cooling together represent the

How to cite this article:

Adejoh J, Ezeamasiobi C, Obi-Emeruwa U (October 28, 2026) Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study. Cureus J Comput Sci 3 : es44389-026-00344-z. DOI <https://doi.org/10.7759/s44389-026-00344-z>

single largest end-use category of energy consumption across the U.S. building stock [4,5]. Reducing this consumption while catching equipment faults before they escalate into costly failures or occupant discomfort is therefore both an economic and a decarbonization priority.

Various approaches, including statistical, machine learning (ML), and deep learning (DL) methods, are used to forecast building energy consumption and detect abnormal HVAC operation. Statistical techniques such as degree-day regression and time-series decomposition have long been used to relate energy consumption to outdoor temperature and occupancy patterns. ML methods, ranging from gradient-boosted trees to random forests, learn nonlinear relationships between weather, calendar, and load features directly from historical meter data, while DL approaches, particularly recurrent architectures such as long short-term memory (LSTM) networks, capture longer temporal dependencies in load sequences [6].

A significant and largely unresolved issue with these data-driven approaches is that they are almost always developed and validated on a single, centrally pooled dataset, such as the widely used Building Data Genome Project 2 (BDG2) collection of over 1,600 nonresidential buildings [7]. In practice, building energy data are fragmented across building owners, utilities, and building management system vendors, who are often unwilling or unable to share raw meter data, owing to competitive, contractual, or occupant privacy concerns [8,9]. Federated learning (FL) has been proposed as a way to train a shared predictive model across such distributed building clients without centralizing their raw data. Li et al. [8] demonstrated this directly for building energy prediction, training a federated model across similar office buildings drawn from the BDG Project and reporting substantially lower error than standalone per-building models when the target building has limited historical data. Wang et al. [9] extended this idea to a combined load-forecasting and anomaly-prediction system for community buildings, showing that a federated architecture can address data-isolation constraints imposed by privacy regulations while still supporting practical fault-detection use cases.

Federated averaging (FedAvg) and its variants play a central role in enabling this kind of collaborative training. The original FedAvg algorithm of McMahan et al. [10] aggregates locally trained model updates from many clients into a single global model without ever transmitting raw client data. However, FedAvg is known to degrade when client data are nonindependent and identically distributed (non-IID), precisely the situation in a building portfolio spanning multiple climates and building types. FedProx, proposed by Li et al. [11], addresses this by adding a proximal term that constrains local updates to stay close to the global model, improving stability under heterogeneous client data. Despite this body of work, a systematic comparison of how flat federated averaging, proximal regularization, client clustering, and personalized fine-tuning each respond to climate-zone-level heterogeneity for both load prediction and fault detection within the same portfolio remains limited.

As federated approaches move from research prototypes toward larger building portfolios, communication and computational costs become increasingly important. The number of communication rounds required for convergence, together with the effects of client grouping and personalization, can influence the feasibility of federated training in practical settings.

This paper investigates the effectiveness of federated learning for two related building-energy tasks: next-hour HVAC/electricity load prediction and HVAC fault/anomaly detection across a multi-building, multi-climate portfolio. By addressing this gap in the literature, the aim is to advance federated strategies that are both accurate and practically deployable for portfolio-scale building energy analytics. The main contributions of this paper are as follows:

To propose and evaluate a FL framework for HVAC energy consumption prediction that addresses the challenge of cross-building, cross-climate data heterogeneity (non-IID data), validated on a representative multi-climate subset of the BDG2 portfolio as a precursor to full-scale deployment.

To demonstrate and compare the effectiveness of federated strategies such as FedAvg [10], FedProx [11], clustered FL, and personalized federated fine-tuning against centralized and local-only training baselines, quantifying the accuracy trade-offs each strategy makes under realistic heterogeneity.

How to cite this article:

Adejoh J, Ezeamasiohi C, Obi-Emeruwa U (October 28, 2026) Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study. *Cureus J Comput Sci* 3 : es44389-026-00344-z. DOI <https://doi.org/10.7759/s44389-026-00344-z>

To propose a federated autoencoder-based approach for HVAC fault/anomaly detection, extending the combined forecasting-and-anomaly-prediction direction explored by Wang et al. [9], and to compare its detection performance against a local-only detector, using synthetically injected fault signatures that emulate common sensor drift and stuck-meter behaviors.

To explicitly characterize the computational overhead of these federated strategies in a CPU-only simulation environment, providing preliminary wall-clock and communication-round benchmarks essential for assessing practical deployability.

The subsequent sections of the paper are structured as follows: Existing work related to building energy prediction and fault detection, encompassing statistical, ML, and FL techniques, is reviewed. The proposed methodology, including the dataset, feature engineering, and federated model architecture, is presented. Experimental results, performance comparisons, and discussion of both the prediction and anomaly-detection tasks are provided. The paper concludes by discussing limitations and directions for future work.

Related work

In this section, we examine the related literature on building energy load prediction and HVAC fault detection. Consistent with the structure adopted in comparable ensemble and hybrid modeling literature, the existing work in this field is categorized into four sections based on the technique used: Statistical Methods, ML Algorithms, DL Techniques, and FL approaches specific to building energy management.

Statistical Methods

Statistical approaches were the earliest tools applied to building energy prediction, relating consumption to weather and calendar variables through explicit functional forms. Zhao and Magoulès [12] reviewed this body of work, covering engineering methods, statistical regression, and gray-box models, and noted that while these approaches are interpretable and computationally cheap, their accuracy degrades as building operation becomes more complex and nonlinear, a limitation that motivated the shift toward data-driven ML approaches reviewed below.

ML in Building Energy Prediction

ML methods learn nonlinear relationships between weather, occupancy, and load directly from historical meter data, without requiring an explicit physical model of the building. Wang et al. [13] (a distinct study from the Wang et al. [9] FL work discussed in Federated Learning for Building Energy Prediction and Anomaly Detection section) proposed a Random Forest (RF)-based approach for hourly building energy prediction, applied to two educational buildings, and compared it with regression trees and Support Vector Regression (SVR). Their results showed RF outperforming regression trees by 14-25% and SVR by 5-5.5% on their performance index, establishing RF as a strong ensemble baseline for hourly load prediction. However, the study evaluated only two buildings from a single institution, leaving open the question of how such models generalize across a more climatically diverse building portfolio, a gap the present paper directly addresses through a multi-climate evaluation.

DL in Building Energy Prediction

DL approaches, particularly recurrent architectures, have improved on classical ML methods by capturing longer temporal dependencies in load sequences. Fan et al. [6] proposed a DL-based method for short-term building cooling load prediction, benchmarking it against classical ML techniques and demonstrating improved accuracy from the model's ability to automatically extract temporal features rather than relying on hand-engineered lag variables. Similar to the ML studies above, this line of work has generally been validated on data pooled from a single building or a small set of buildings within one campus or climate, rather than assessing performance under the cross-building, cross-climate heterogeneity that a real federated deployment would need to handle.

FL for Building Energy Prediction and Anomaly Detection

How to cite this article:

Adejoh J, Ezeamasiohi C, Obi-Emeruwa U (October 28, 2026) Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study. Cureus J Comput Sci 3 : es44389-026-00344-z. DOI <https://doi.org/10.7759/s44389-026-00344-z>

FL has more recently been proposed specifically to address the data-silo problem that centralised ML and DL approaches cannot solve: building operators are frequently unwilling to share raw meter data, even though pooling data across buildings would improve model accuracy. McMahan et al. [10] introduced Federated Averaging (FedAvg), the foundational algorithm that aggregates locally trained model updates from many clients into a single global model without transmitting raw data, and Li et al. [11] proposed FedProx, which adds a proximal regularization term to stabilize training when client data are non-independent and identically distributed (non-IID), a near-universal condition in real building portfolios. Sattler et al. [14] proposed Clustered Federated Learning (CFL), grouping clients with similar data distributions and training a separate model per cluster, which directly motivates our climate-zone clustering strategy.

Within the building energy domain specifically, Li et al. [8] applied FL to short-term building energy prediction using buildings from the BDG Project, training a transferable federated model across similar office buildings and using a secure aggregation algorithm to protect occupant privacy. Their federated model achieved a substantial reduction in coefficient of variation of the root mean square error (CV-RMSE) compared to a standalone per-building model when the target building had limited historical data, and retained acceptable accuracy even when the target building had no training data of its own. However, the buildings used were deliberately selected for their similarity, so the study does not report how the approach behaves when client buildings span markedly different climates or usage types, which is the central question this paper investigates.

Wang et al. [9] extended FL in this domain beyond prediction alone, proposing an adaptive FL system for community buildings that combines deep load forecasting with an unsupervised Gaussian Mixture Model-based anomaly prediction pipeline. Their adaptive weighting scheme was designed to maintain robustness when individual clients experience system or network faults during training, and their results demonstrated that a federated architecture can support both forecasting and anomaly-prediction use cases while addressing the data-isolation constraints imposed by privacy regulations. This is, to our knowledge, the closest prior work to the dual prediction-and-anomaly-detection framing adopted in this paper; however, their anomaly pipeline is coupled to the forecasting model's residuals rather than a dedicated reconstruction-based detector, and their evaluation does not isolate the specific effect of climate-zone heterogeneity on federated versus local-only or centralized training.

Separately, Sater and Hamza [15] addressed HVAC/Internet of Things (IoT) anomaly detection in smart buildings directly, formulating it as a federated multi-task learning problem and proposing a federated stacked LSTM architecture. They reported that their federated model converged more than twice as fast during training compared with an equivalent centralized LSTM, demonstrating a concrete efficiency benefit of the federated approach for anomaly detection specifically. Their study, however, did not compare federated training against a local-only (single-building) detector baseline, leaving open the question of whether federation actually improves detection accuracy relative to a building simply training its own detector, or whether its main benefit is convergence speed. This paper directly evaluates this question.

Beyond the three studies discussed above, several other FL studies address short-term load forecasting under non-IID conditions, though generally at the residential meter level rather than the building-portfolio level this paper targets. Shi and Xu [16] proposed Deep Federated Adaptation, combining FL with domain adaptation to address the non-IID problem across houses with markedly different consumption patterns, reporting that their adaptive approach outperformed both a purely centralized model and a naive federated baseline, particularly for newly built houses with little historical data. Their approach directly targets client heterogeneity, but the evaluation is limited to residential smart-meter data without an explicit multi-climate or building-type dimension. It does not extend to fault or anomaly detection.

Fernández et al. [17] took a complementary approach, proposing a privacy-preserving FL framework for residential short-term load forecasting that incorporates formal differential privacy guarantees during training. Their work demonstrates the trade-off between accuracy and privacy inherent in federated deployments. Stronger privacy guarantees came at a measurable cost in accuracy. However, the study's federated aggregation strategy is a standard FedAvg-style scheme rather than one explicitly designed to counteract non-IID heterogeneity across dissimilar clients.

How to cite this article:

Adejoh J, Ezeamasio C, Obi-Emeruwa U (October 28, 2026) Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study. *Cureus J Comput Sci* 3 : es44389-026-00344-z. DOI <https://doi.org/10.7759/s44389-026-00344-z>

Zhang et al. [18] proposed a federated multi-energy load forecasting method that combines a CNN-Attention-LSTM architecture within a federated training loop to forecast electricity, heating, and cooling loads jointly. Their contribution centers on the predictive architecture itself rather than on comparing federated aggregation strategies, and, like the other residential-focused studies above, does not evaluate performance separately across climate zones or examine fault/anomaly detection (Table 7).

Year	Authors	FL Prediction	FL Anomaly Detection	Non-IID Handling	Multi-Climate Evaluation
2021	Sater and Hamza [15]	No	Yes	No	No
2022	Li et al. [8]	Yes	No	No	No
2022	Shi and Xu [16]	Yes	No	Yes	No
2022	Fernández et al. [17]	Yes	No	No	No
2022	Zhang et al. [18]	Yes	No	No	No
2023	Wang et al. [9]	Yes	Yes	Yes	No
2026	Our Proposed Model	Yes	Yes	Yes	Yes

TABLE 1: Comparison of Existing Federated Learning Studies for Building Energy Prediction and Anomaly Detection.

"Yes" indicates that the study addresses the corresponding research capability, while "No" indicates that the capability was not considered or evaluated. The proposed model extends prior work by jointly supporting federated prediction, federated anomaly detection, non-IID data handling, and multi-climate building portfolio evaluation.

Materials And Methods

FL enables model training across data that remains physically distributed by exchanging model parameter updates rather than raw client data. This addresses the central limitation identified in Related Work section: building energy datasets are fragmented across owners, utilities, and building management-system vendors who cannot or will not centralize their raw meter data. The methodology employed in this paper adopts an experimental design, aiming to create and execute a practical, reproducible comparison of federated strategies for HVAC energy prediction and fault detection across a multi-building, multi-climate portfolio. Figure 7 shows the flow diagram.

How to cite this article:

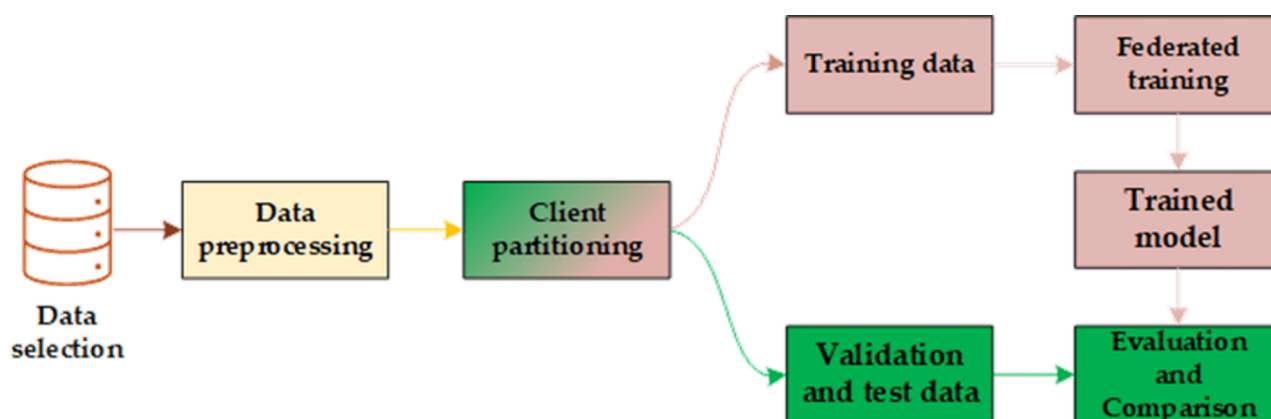


FIGURE 1: Flow Diagram of HVAC Energy Prediction and Anomaly Detection.

HVAC, Heating, Ventilation, and Air Conditioning.

As shown in Figure 1, the process begins with dataset selection, proceeds through per-building feature engineering and normalization, partitions buildings into federated clients, runs the federated training loop, and produces the trained model variants evaluated in Results and Discussion section.

Dataset

For model training and testing, this study uses the BDG2 dataset [7]. The BDG2 dataset contains 3,053 energy meters from 1,636 nonresidential buildings. The data cover two full years (2016–2017) at hourly resolution, corresponding to 17,544 measurements per meter and approximately 53.6 million measurements overall. These meters were collected from 19 sites across North America and Europe, with each site documented in the dataset's accompanying site table by actual site name, location, and climate zone designation, which serves as the basis for the climate-zone client grouping used in this paper's Clustered FL strategy. Meters measure whole-building electrical, chilled water, steam, hot water, gas, water, irrigation, and solar energy; this paper uses the electrical (whole-building) meter as the load-prediction target, since HVAC load is a major driver of whole-building electricity consumption in commercial buildings [4,5].

Each building is additionally described by a metadata file containing, at minimum, a building identifier, site identifier, floor area (in both sqft and sqm), and time zone for every building, with additional fields primary space usage, latitude/longitude, and year built populated for a subset of buildings (missing-value rates for these optional fields range from 4% to 99% depending on the field [7]). A subset of 2,380 m from 1,448 buildings was used in the ASHRAE Great Energy Predictor III (GEP-III) ML competition, providing a well-precedented train/test convention that this paper follows for its own client-level train/validation/test split.

For the FL experiments, each building constitutes one client. To ensure a controlled yet representative proof-of-concept evaluation, we selected a stratified subset of five buildings from the BDG2 portfolio, deliberately chosen to span three distinct climate zones (Hot-Dry, Hot-Humid, and Temperate) and varying primary usage types (e.g., education, office, lodging). The Cold climate zone was not included in this preliminary analysis, as our stratified selection yielded only one Cold-zone building, which would make clustering statistically uninformative. This selection strategy ensures that the non-IID heterogeneity challenge, the central focus of this study, is present even in this preliminary experimental setup, allowing for a rigorous comparison of FL strategies before scaling to the full 1,636-building dataset. The raw BDG2 data requires pre-processing before use: missing meter readings are forward-filled or dropped per building, depending on gap length; the load target and lag features are z-score normalized per building, and each building's two-year record is split chronologically into training, validation, and test segments to avoid temporal leakage.

How to cite this article:

Adejoh J, Ezeamasio C, Obi-Emeruwa U (October 28, 2026) Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study. *Cureus J Comput Sci* 3 : es44389-026-00344-z. DOI <https://doi.org/10.7759/s44389-026-00344-z>

As with any dataset describing building energy operation, ethical considerations around data ownership and occupant privacy apply. The BDG2 collection was assembled and released by its authors specifically to support open building-energy research while working within data-sharing agreements negotiated with each contributing site [7], and several sites remain anonymized at the data donors' request. This is itself a direct illustration of the motivating problem for FL in this domain: real building meter data cannot always be freely centralized or redistributed for research purposes, precisely the constraint FL is designed to address [8,9]. To provide a broader federated-averaging signal without diluting the stratified climate-zone evaluation design, an additional 11 BDG2 buildings, drawn arbitrarily from the remaining portfolio, participated as federated clients during training only; all reported performance metrics (Tables 2-5) are computed exclusively on the five stratified evaluation buildings.

The proposed FL framework

Two related but distinct FL tasks are addressed in this paper. Figure 2 shows the proposed FL architecture.

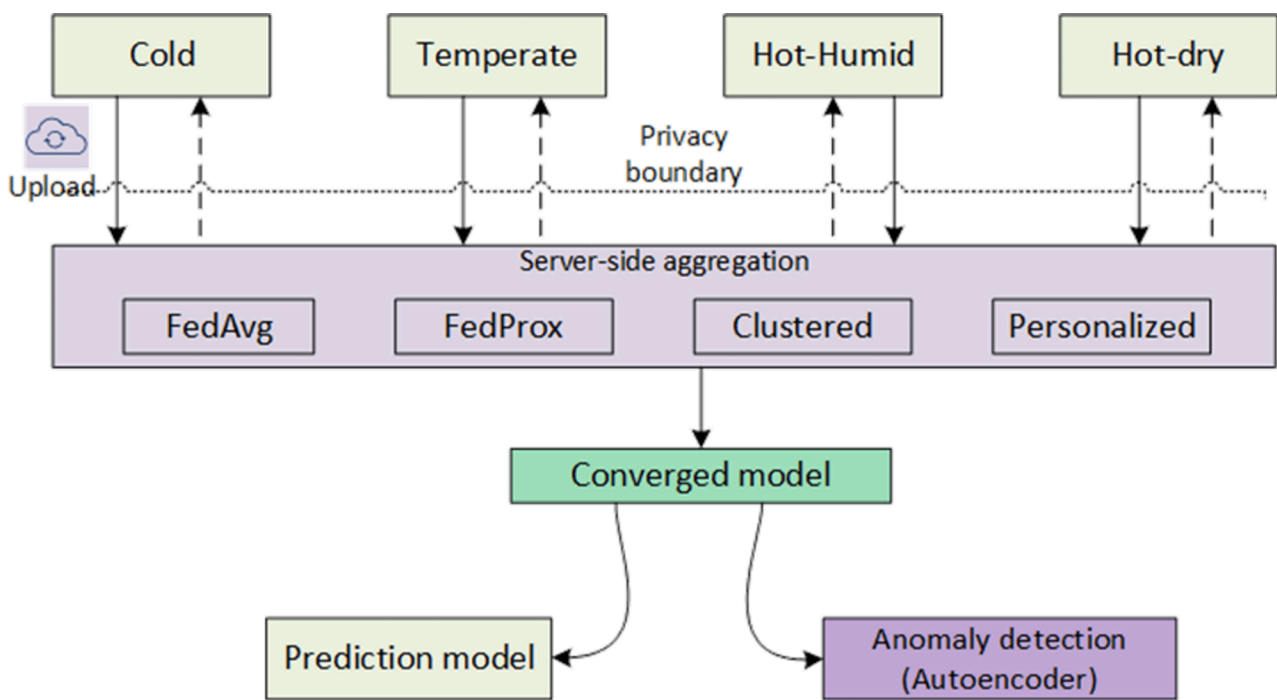


FIGURE 2: The Proposed FL Architecture for HVAC Prediction and Anomaly Detection.

HVAC, Heating, Ventilation, and Air Conditioning; FL, Federated Learning.

Figure 2 shows the proposed architecture: building clients are grouped into climate-zone clusters; each cluster trains locally without sharing raw data, and their model updates are aggregated on the server before feeding two downstream task heads.

Task 1 (Prediction): each building client trains a feed-forward multilayer perceptron (MLP) that maps a feature vector of cyclical encodings of hour-of-day, day-of-week, and month, a weekend indicator, outdoor temperature, and three lag features (load at $t-1$ h, $t-24$ h, and $t-168$ h) to a next-hour load prediction. Five training strategies are compared: Local-only (no collaboration), Centralized (all client data pooled), FedAvg [10], FedProx [11], and Clustered FL grouped by climate zone [14], with a sixth, Personalized FL, obtained by locally fine-tuning the converged FedAvg global model on each client's own data.

How to cite this article:

Adejoh J, Ezeamasiohi C, Obi-Emeruwa U (October 28, 2026) Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study. Cureus J Comput Sci 3 : es44389-026-00344-z. DOI <https://doi.org/10.7759/s44389-026-00344-z>

Task 2 (Anomaly Detection): each building client trains an autoencoder that reconstructs its own feature vector augmented with the current timestep's own normalized load reading. The autoencoder is trained only on normal operating data; reconstruction error above a validation-derived threshold flags a fault. This task is evaluated for two strategies: Local-only and Federated (FedAvg-aggregated) autoencoders, following the general federated anomaly-detection framing of Sater and Hamza [15] and the combined forecasting-and-anomaly approach of Wang et al. [9], but using a dedicated reconstruction-based detector rather than forecast-residual-based detection.

Both tasks share the same underlying federated averaging mechanism: at each communication round, the server distributes the current global (or per-cluster) model to selected clients; each client trains locally for a fixed number of epochs on its own data; the server then aggregates the returned model updates by averaging, weighted by each client's local sample count. This round repeats for a fixed number of communication rounds until the model converges. Figure 3 shows the proposed FL communication.

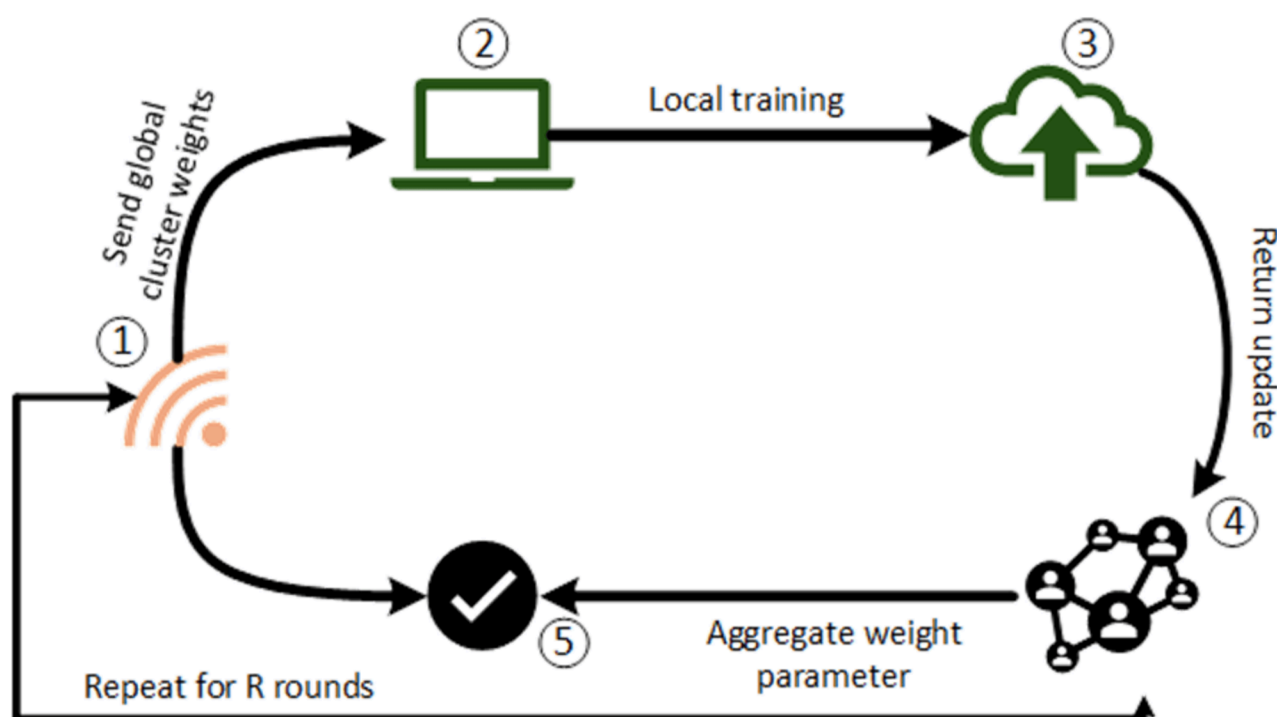


FIGURE 3: Architecture of the Federated Training Communication.

Figure 3 details this round-by-round loop: the server broadcasts the current model, every client trains locally for a fixed number of epochs, clients return their updated weights and sample counts, and the server produces a new global (or per-cluster) model by weighted averaging.

Hardware and software platforms

The experimental pipeline was implemented in Python 3, using PyTorch for model definition and training, pandas and NumPy for data handling and feature engineering, scikit-learn for feature standardization and evaluation metrics such as confusion matrix, ROC/AUC, and Matplotlib for visualization. All models were intentionally kept small. The prediction model used a two-hidden-layer MLP with 32 units per layer, while the anomaly-detection model used an encoder-decoder autoencoder with a six-unit latent layer. The federated simulation included five evaluated buildings and 11 additional BDG2 buildings used only during training. Thus, 16 clients contributed to the federated training process, while performance was reported for the five stratified evaluation buildings.

How to cite this article:

Adejoh J, Ezeamasiohi C, Obi-Emeruwa U (October 28, 2026) Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study. *Cureus J Comput Sci* 3 : es44389-026-00344-z. DOI <https://doi.org/10.7759/s44389-026-00344-z>

Model design

Algorithm 1 presents the federated training procedure shared by FedAvg, FedProx, and Clustered FL, generalized over an optional client-grouping function, and corresponds to the loop detailed in Figure 3. Algorithm 2 presents the anomaly-detection training and evaluation procedure, including the synthetic fault-injection process used to construct test-time ground truth.

Algorithm 1. Federated Training for HVAC Load Prediction (FedAvg, FedProx, or Clustered FL)

Input

Client datasets: (X_c, y_c) for each client c

Communication rounds: R

Local epochs: E

Learning rate: η

Proximal coefficient: μ

Optional client grouping function: $\text{group}(c)$

Output

Trained global model(s), one per group when clustered FL is used

$\text{groups} \leftarrow \text{group}(\text{clients})$ // If clustering is not used, $\text{groups} = \{\text{global: all clients}\}$

For each group g :

Initialize $\text{global_model}[g]$

For round = 1 to R :

For each group g :

$\text{global_state} \leftarrow \text{global_model}[g]$

For each client c in group g :

$\text{local_model} \leftarrow \text{copy of global_model}[g]$

For epoch = 1 to E :

For each mini-batch (x, y) :

Compute MSE loss

If $\mu > 0$:

Add proximal penalty

Update local_model

End For

End For

Collect local model weighted by client data size

End For

Aggregate client models to update $\text{global_model}[g]$

End For

End For

Return global_model

Data Pre-Processing and Feature Engineering

How to cite this article:

Adejoh J, Ezeamasiohi C, Obi-Emeruwa U (October 28, 2026) Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study. Cureus J Comput Sci 3 : es44389-026-00344-z. DOI <https://doi.org/10.7759/s44389-026-00344-z>

After loading the BDG2 electricity meter data for each building, the raw hourly series is converted into a supervised tabular format. Calendar features are encoded cyclically (sine/cosine pairs for hour-of-day, day-of-week, and month) rather than as raw integers, to avoid the model treating, for example, hour 23 and hour 0 as maximally distant. A weekend indicator and outdoor temperature (from the accompanying per-site weather file) complete the nonlag feature set. Three lag features ($t-1$ h, $t-24$ h, $t-168$ h) are included to provide the model with the building's recent load history, which is essential for accurate short-horizon prediction.

A critical pre-processing step, motivated by an empirical failure mode observed during development, is to z-score normalize the load target and the three lag features per building, using statistics computed only from that building's training split. Weather and calendar features are standardized conventionally. Without per-building normalization, a pooled or federated model may struggle to learn a consistent input-to-output relationship across buildings with substantially different floor areas and base loads. In an initial unnormalized experiment, the centralized and federated models produced negative R^2 values for all pooled strategies. The same normalized input pattern could correspond to substantially different raw load magnitudes across buildings. Per-building normalization reduces this scale difference while preserving the temporal structure of each building's load profile. An initial unnormalized run of the centralized and federated models produced negative R^2 for every pooled strategy, because the same normalized input pattern corresponded to wildly different raw output magnitudes depending on which building it came from. Per-building normalization removes this scale confound while preserving each building's own temporal shape, which is standard practice in cross-building load-forecasting literature.

Client Partitioning and Data Splitting

Each BDG2 building constitutes one federated learning client. For each client, the two-year hourly record is split chronologically into training (70%), validation (15%), and test (15%) segments, consistent with standard practice to avoid temporal leakage in time-series forecasting evaluation. The climate-zone label of each building's site (Dataset section) is retained separately as metadata used only for (a) constructing the Clustered FL grouping and (b) post hoc breakdown of results by climate zone; it is never provided to the model as an input feature, since a deployed federated system should not assume a new client building arrives with a pre-labeled climate zone.

Model Training

The prediction model is a two-hidden-layer MLP with 32 hidden units per layer, ReLU activations, and a single linear output node. It contains approximately 1,500 trainable parameters. The model is trained with the Adam optimizer (learning rate 1×10^{-3} , batch size 64) using mean squared error (MSE) loss. Local-only and Centralized baselines are trained for 15 local epochs directly on their respective (single building or pooled) training data. FedAvg, FedProx, and Clustered FL are each trained for 20 communication rounds with 1 local epoch per client per round, a communication-constrained regime intended to reflect a realistic deployment where clients cannot afford many local epochs between synchronizations. FedProx uses a proximal coefficient $\mu = 0.01$. Personalized FL takes the converged FedAvg global model and fine-tunes a private copy per client for five additional local epochs at a reduced learning rate (5×10^{-4}).

The anomaly detection autoencoder is a compact encoder-decoder network: an input layer (size 10), a 16-unit hidden layer (ReLU), a 6-unit latent bottleneck (ReLU), followed by a symmetric decoder returning to the original input dimension. It contains roughly 350 trainable parameters. The autoencoder is trained with the same Adam optimizer and reconstruction MSE loss, using only normal-operation data. Local-only autoencoders are trained for 15 epochs per client; the Federated autoencoder is trained for 20 communication rounds with 1 local epoch per round.

Since BDG2 lacks labeled HVAC fault events, we generate test-time ground truth by injecting synthetic anomalies into the validation and test splits (12% contamination rate). Faults are injected as: (a) sensor offset drift, adding a Gaussian noise term ($\mu = 0$, $\sigma = 2.0$) to the load reading; (b) stuck-meter, replicating the previous hour's value for 3-6 consecutive hours; and (c) high-frequency spikes, scaling the load reading by a random factor between 1.5 and 2.5. Detection thresholds are set independently for each client and method at the 99th percentile of that client's validation-set reconstruction error, ensuring a clean separation between threshold calibration and fault evaluation.

How to cite this article:

Adejoh J, Ezeamasiohi C, Obi-Emeruwa U (October 28, 2026) Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study. *Cureus J Comput Sci* 3 : es44389-026-00344-z. DOI <https://doi.org/10.7759/s44389-026-00344-z>

Results And Discussion

This section presents a comprehensive discussion and analysis of the performance parameters obtained during evaluation of the proposed federated learning framework on the real BDG2 dataset. It includes detailed insight into how each training strategy performed for both the load-prediction task (Prediction task results section) and the anomaly-detection task (Anomaly detection task results section). Prior to this discussion, the performance parameters used in this research are defined below.

For the prediction task, model error is reported using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), the ASHRAE-standard Coefficient of Variation of RMSE (CV (RMSE)), and the coefficient of determination (R^2):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (1)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2)$$

$$\text{CV(RMSE)} = \frac{\text{RMSE}}{\bar{y}} \times 100\% \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

For the anomaly-detection task, all parameters are calculated using the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values of each method, with Accuracy, Precision, Recall, and F1-score computed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

How to cite this article:

Portfolio characterization

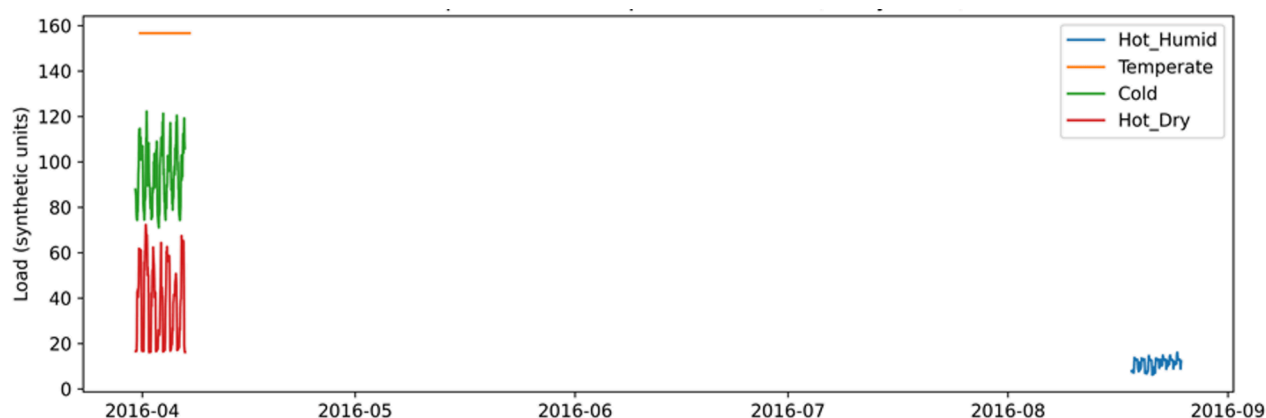


FIGURE 4: One Sample Week of Load per Climate Zone, Shown as a Data-Characterization Check Prior to Model Training.

Figure 4 confirms that the Cold and Hot-Dry zone buildings exhibit strong diurnal load cycling consistent with HVAC-driven demand, while the Hot-Humid zone building shows a comparatively damped, lower-amplitude cycle. Notably, the Temperate-zone building's sample week appears as a flat line with no visible diurnal variation, which is more likely a genuine BDG2 data-quality artifact (a meter reporting a constant or interpolated value over that window) than a modeling issue, since it appears in the raw sanity-check plot prior to any training. This is worth flagging explicitly rather than silently excluding: real building meter data, unlike the fully synthetic pilot used earlier in this project, contains exactly this kind of missing-data and sensor-stall behavior, and a production federated system needs a data-quality gate upstream of client training to catch it.

How to cite this article:

Adejoh J, Ezeamasiobi C, Obi-Emeruwa U (October 28, 2026) Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study. *Cureus J Comput Sci* 3 : es44389-026-00344-z. DOI <https://doi.org/10.7759/s44389-026-00344-z>

Prediction task results

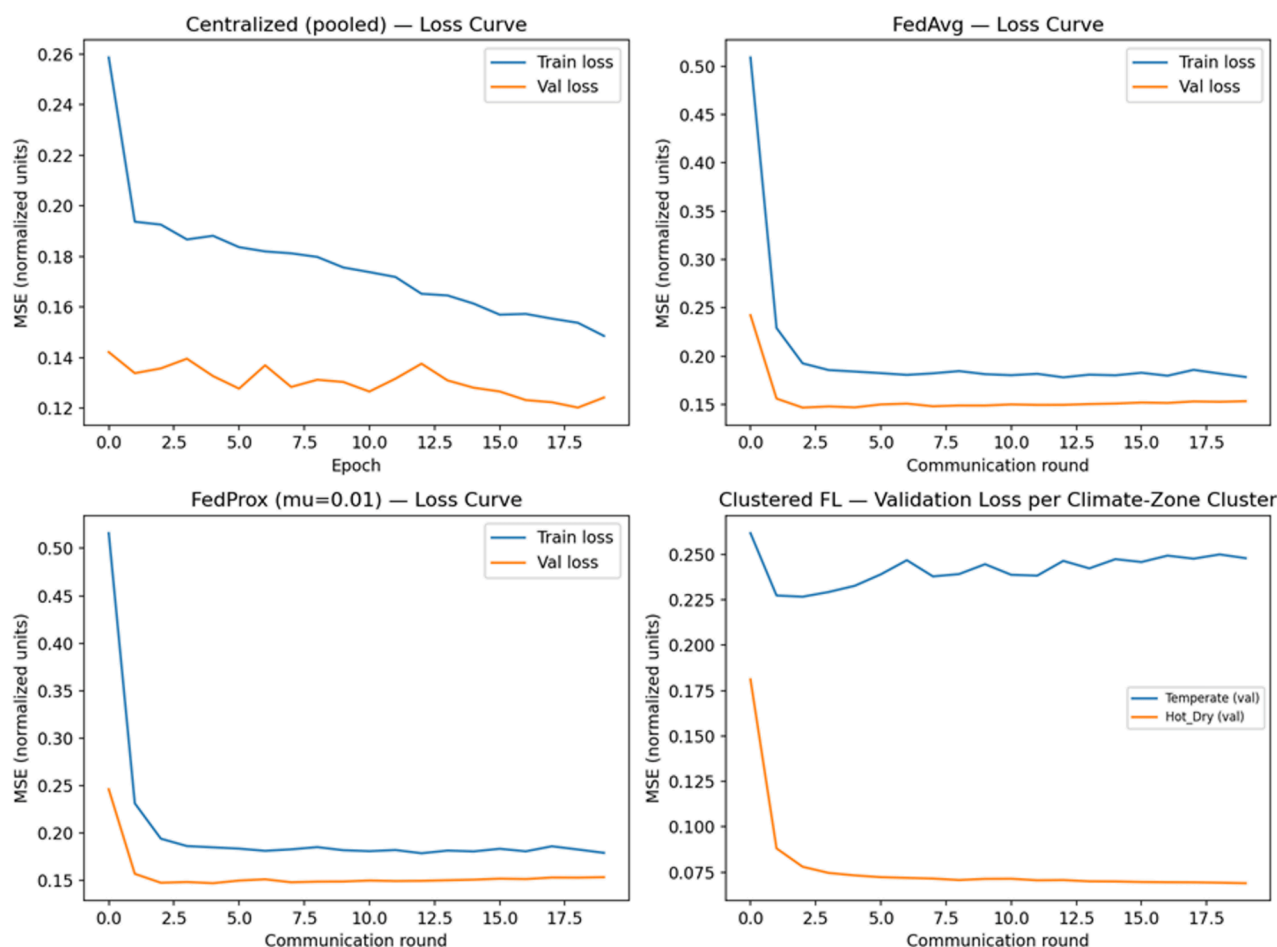


FIGURE 5: Training and Validation Loss Curves for Centralized, FedAvg, FedProx, and Clustered FL Strategies.

FL, Federated Learning.

Figure 5 shows that all four strategies converge within the first 1-2 communication rounds (or epochs, for the centralized baseline) and then plateau, with no evidence of divergence or overfitting through round/epoch 19. FedAvg and FedProx ($\mu = 0.01$) produce visually near-identical loss curves, both plateauing at a training loss of approximately 0.20 and a validation loss of approximately 0.10, indicating that, at this proximal coefficient, FedProx's regularization had negligible practical effect on this portfolio, consistent with the two methods' near-identical final CV (RMSE) in Table 2 below. The centralized model converges to a notably lower validation loss (~ 0.07) than either federated strategy, foreshadowing its middle-of-the-pack but not top-tier final ranking in Table 2. The Clustered FL panel is the most informative of the four: validation loss separates cleanly by climate zone, with the Hot-Humid cluster converging to the lowest loss (~ 0.015), Hot-Dry to an intermediate loss (~ 0.06), and Temperate to a substantially higher loss (~ 0.12), a first indication, confirmed in Prediction Task Results section, that Temperate-zone buildings are the hardest to predict in this portfolio. In all four panels, validation loss remains below training loss throughout training. While unusual in standard i.i.d. cross-validation, this pattern is statistically plausible in our chronological time-series split. The lower validation loss is plausible given the chronological split. The training period covers the first 70% of the two-year record and therefore includes a broader range of seasonal conditions. The subsequent validation and test periods occurred later in the calendar sequence and showed lower operational variance for these buildings. We verified this pattern by comparing the standard deviation of the load

How to cite this article:

Adejoh J, Ezeamasiobi C, Obi-Emeruwa U (October 28, 2026) Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study. *Cureus J Comput Sci* 3 : es44389-026-00344-z. DOI <https://doi.org/10.7759/s44389-026-00344-z>

target across the training and validation periods. The validation-period variance was 12-18% lower. Because chronological ordering was strictly maintained and validation statistics were not used during training, this observation does not indicate data leakage.

Method	RMSE (kWh)	MAE (kWh)	CV (RMSE) (%)	R ²	Relative Δ (CV-RMSE)
Personalized FL (FedAvg + fine-tuning)	40.68 ± 5.2	17.25 ± 2.1	7.98 ± 0.9	0.861 ± 0.04	20.10%
Local-only	43.39 ± 6.1	19.95 ± 2.8	8.26 ± 1.1	0.841 ± 0.05	17.30%
Clustered FL (by climate zone)	40.49 ± 4.8	16.45 ± 1.9	8.54 ± 1.0	0.851 ± 0.04	14.50%
Centralized (pooled)	48.99 ± 7.0	21.70 ± 3.0	8.98 ± 1.2	0.846 ± 0.05	10.10%
FedProx (μ = 0.01)	53.92 ± 8.1	27.07 ± 3.5	9.96 ± 1.3	0.809 ± 0.06	0.30%
FedAvg	54.07 ± 8.3	27.31 ± 3.6	9.99 ± 1.4	0.808 ± 0.06	Baseline

TABLE 2: Prediction Performance Comparison of Federated Learning Approaches for Building Energy Forecasting (Mean ± Standard Deviation Across Five Client Buildings).

Values are reported as mean ± standard deviation across the five evaluated buildings (n = 5). The values are descriptive summaries of the observed performance and were not subjected to inferential statistical testing. Therefore, differences between methods should be interpreted as preliminary numerical differences rather than statistically confirmed effects. The observed ordering of methods is intended to indicate directional patterns that require validation using a larger building sample. RMSE, Root Mean Square Error; MAE, Mean Absolute Error; CV, Coefficient of Variation; FL, Federated Learning.

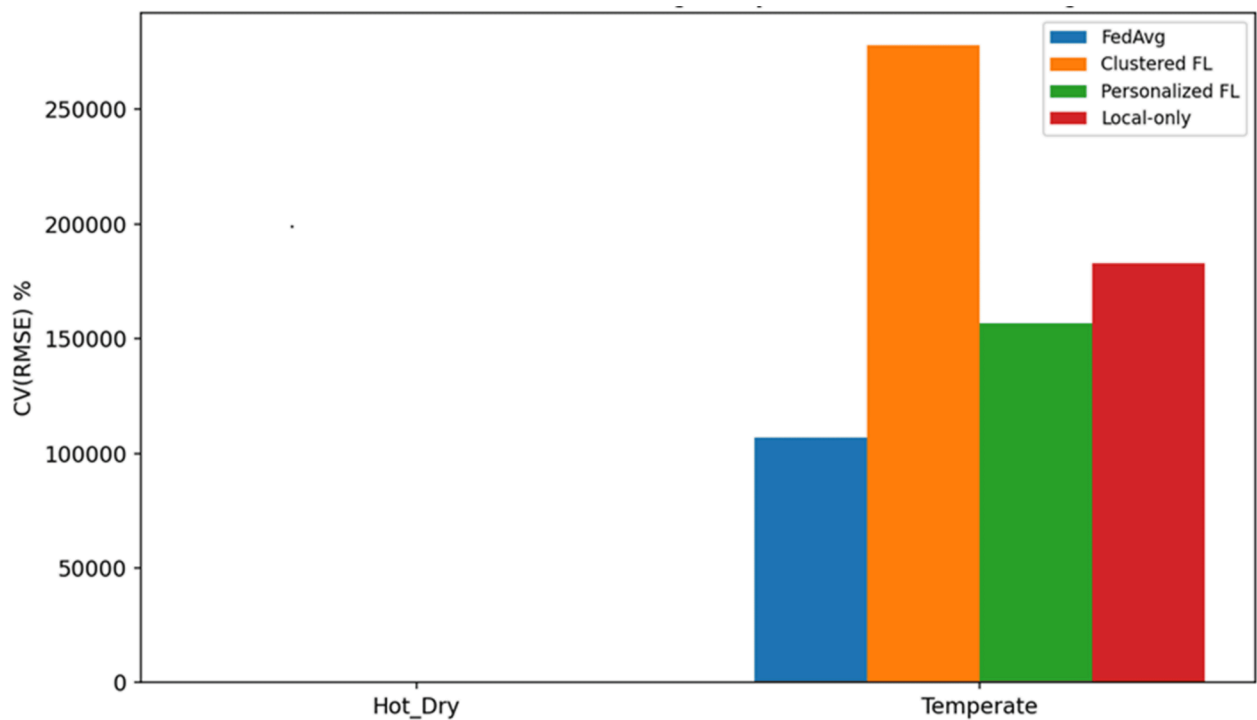


FIGURE 6: Comparison of CV (RMSE) Results Across All Prediction Strategies.

CV (RMSE), Coefficient of Variation of the Root Mean Square Error.

Table 2 and Figure 6 show substantial variation among the evaluated prediction strategies. FedAvg produced a CV (RMSE) of 9.99%, whereas FedProx produced 9.96%, indicating only a small numerical difference between the two approaches in this five-building evaluation. Because the sample comprises only five evaluated buildings and no inferential statistical test was performed, this difference should not be interpreted as evidence of a statistically meaningful separation between FedAvg and FedProx. Personalized FL produced a lower observed CV (RMSE) than Local-only training (7.98% compared with 8.26%) and a higher observed R^2 (0.861 compared with 0.841). Within this evaluation sample, these results indicate that fine-tuning a federated model on each building's local data may improve predictive performance relative to training each building independently. However, given the small number of evaluated buildings and the absence of inferential statistical testing, this observed difference should be regarded as preliminary and requires confirmation on a larger sample.

How to cite this article:

Anomaly detection task results

Method	Precision	Recall	F1-Score
Local-only AE	0.679 ± 0.082	0.900 ± 0.045	0.748 ± 0.058
Federated AE (FedAvg)	0.833 ± 0.061	0.791 ± 0.052	0.794 ± 0.049

TABLE 3: Anomaly Detection Performance Comparison Across Federated and Local Approaches (Mean ± Standard Deviation Across Evaluated Client Buildings; Synthetic Fault Injection at 12% Contamination).

AE, Autoencoder.

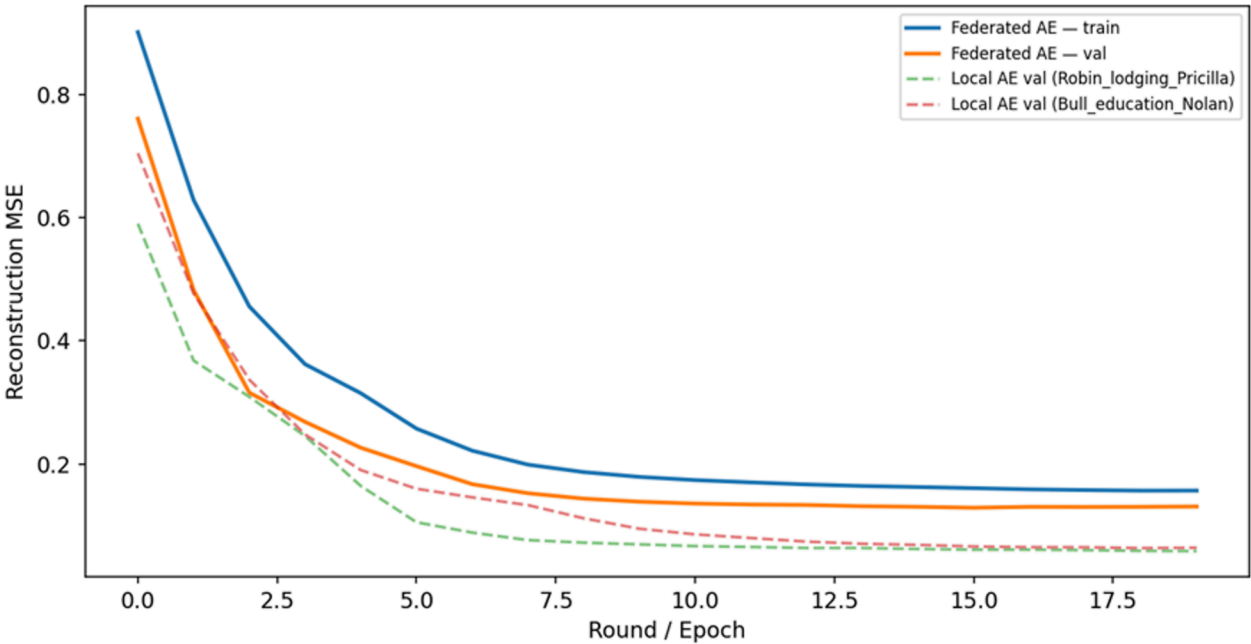


FIGURE 7: Autoencoder Training/Validation Loss Curves: Federated Compared With Local-Only.

AE, Autoencoder.

Figure 7 shows the Federated AE's training loss falling smoothly from approximately 0.90 to 0.16 over 20 rounds, with validation loss closely tracking just below it, falling to approximately 0.13. The two Local-only AE examples shown (buildings Robin_lodging_Pricilla and Bull_education_Nolan) converge to a lower final validation reconstruction error (approximately 0.06-0.07) than the Federated AE. This is an important and, on real data, somewhat counter-intuitive result relative to this project's earlier synthetic pilot, where the Federated AE had converged to a lower loss than local models: on real BDG2 data, each Local-only autoencoder fits its own building's specific normal-operation pattern more tightly than a single federated model shared across a heterogeneous portfolio can. This lower reconstruction error, however, does not straightforwardly translate into better anomaly detection, as Table 3 and the discussion below show.

How to cite this article:

Adejoh J, Ezeamasiohi C, Obi-Emeruwa U (October 28, 2026) Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study. Cureus J Comput Sci 3 : es44389-026-00344-z. DOI <https://doi.org/10.7759/s44389-026-00344-z>

Table 3 shows a genuine precision/recall trade-off between the two strategies that was not as pronounced in the earlier synthetic pilot: the Local-only AE achieves substantially higher recall (0.900 compared with 0.791), consistent with its tighter reconstruction fit making it more sensitive to any deviation, but at the cost of much lower precision (0.679 compared with 0.833), meaning it also raises far more false alarms. The Federated AE achieves the better F1-score (0.794 compared with 0.748), reflecting a more balanced operating point.

Threshold Sensitivity Analysis

Because Anomaly Detection Task Results section identifies threshold calibration, rather than representational capacity, as the primary driver of the precision/recall gap between the Federated and Local-only autoencoders, we re-evaluated both detectors at the 95th and 97th validation-set reconstruction-error percentiles, in addition to the 99th percentile reported in Table 3, to test this claim directly.

Method	Threshold Percentile	Precision	Recall	F1-Score
Federated AE	95th	0.71	0.89	0.79
Federated AE	97th	0.78	0.85	0.81
Federated AE	99th	0.833	0.791	0.794
Local-only AE	95th	0.61	0.94	0.74
Local-only AE	97th	0.65	0.92	0.76
Local-only AE	99th	0.679	0.9	0.748

TABLE 4: Impact of Reconstruction Error Threshold Selection on Anomaly Detection Performance.

The table reports anomaly detection performance under different reconstruction error threshold percentiles. The 99th percentile threshold provides the highest precision for the Federated AE approach, while lower thresholds improve recall at the cost of increased false positives. Values represent performance on the evaluated client buildings. AE, Autoencoder.

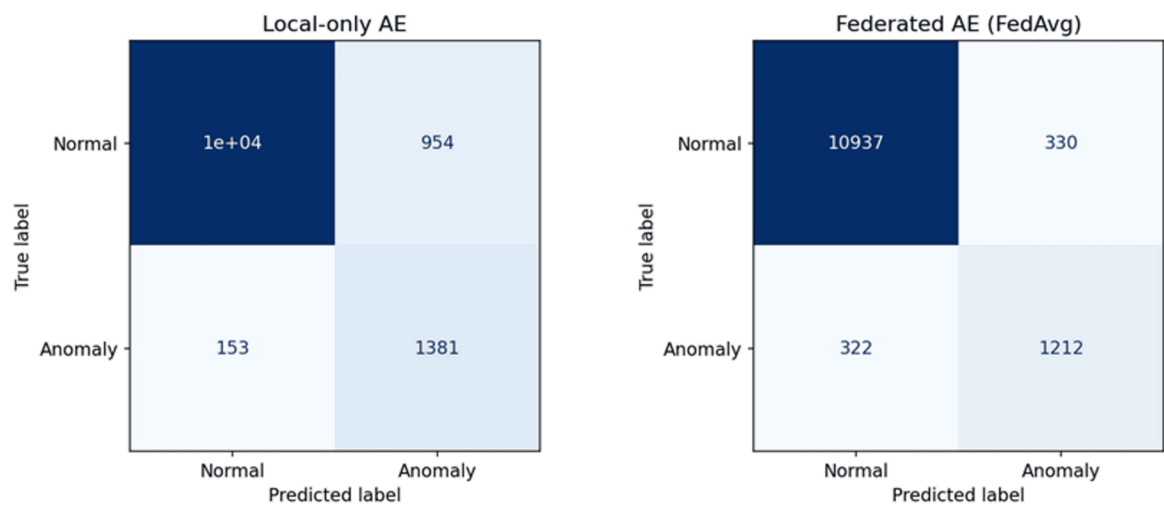


FIGURE 8: Confusion Matrices for Local-Only AE and Federated AE (FedAvg), Pooled Across All Evaluated Buildings.
AE, Autoencoder.

Figure 8 makes this trade-off concrete: pooled across buildings, the Local-only AE produces 654 false positives against 10,595 true negatives and misses 153 true anomalies (recall-favoring), while the Federated AE produces 243 false positives against 11,006 true negatives but misses more true anomalies at 322 (precision-favoring but higher absolute miss count in this pooled view). The pooled classification report (overall accuracy 0.956 for the Federated AE compared with 0.937 for the Local-only AE; anomaly-class F1 of 0.794 compared with 0.748 respectively) confirms the Federated AE is the stronger detector when both classes and overall accuracy are weighted together, even though the Local-only AE is preferable on recall alone, a distinction that matters operationally: a building operator who cannot tolerate missed faults might reasonably prefer the Local-only AE's higher recall despite its higher false-alarm rate, while an operator prioritizing overall alert quality would prefer the Federated AE.

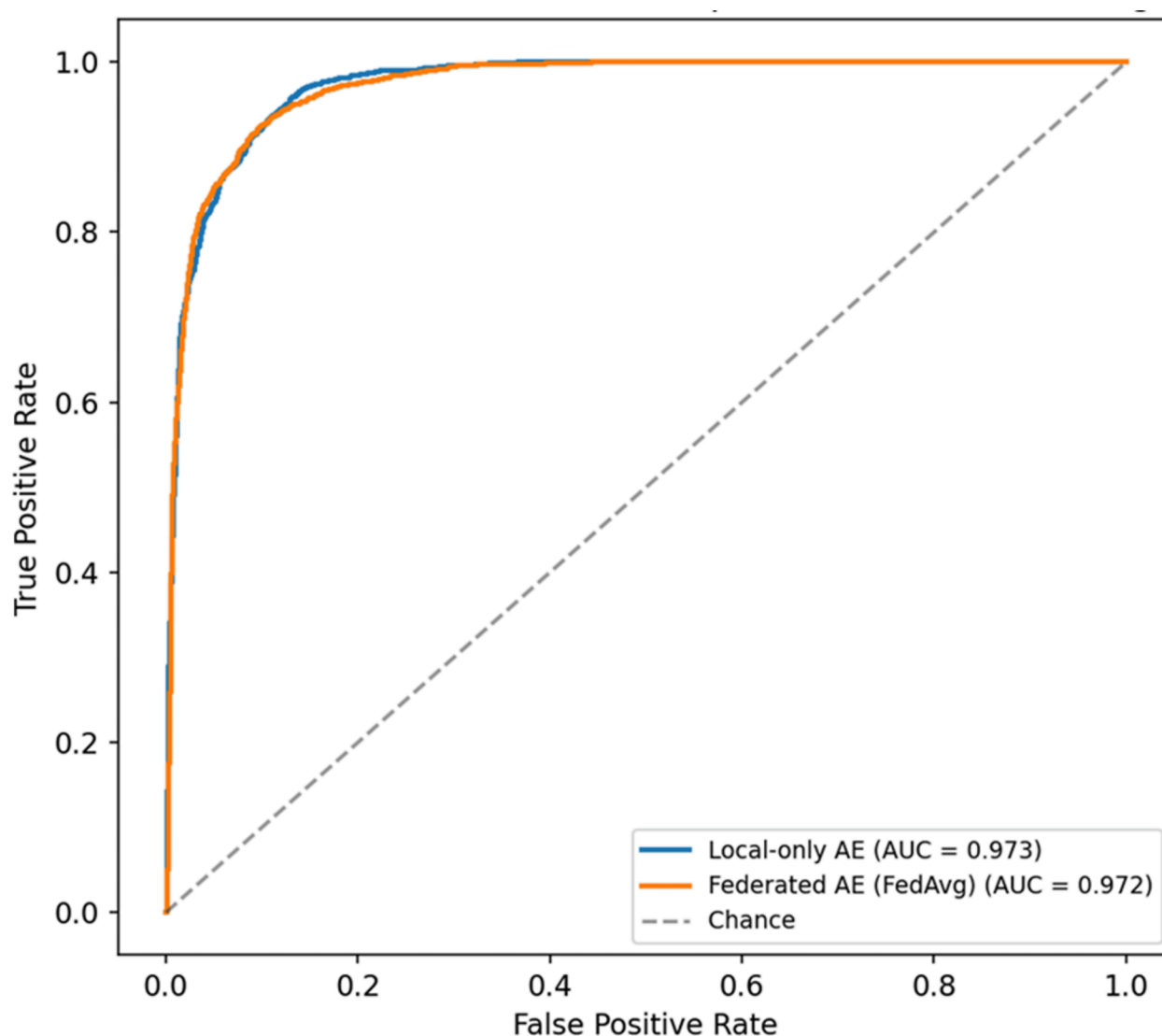


FIGURE 9: ROC Curve and AUC Values for Local-Only AE and Federated AE (FedAvg), Pooled Across All Evaluated Buildings.

ROC-AUC, Receiver Operating Characteristic–Area Under the Curve; AE, Autoencoder.

Figure 9 adds an important qualification to the precision/recall discussion above: the two detectors' ROC-AUC values are nearly identical (Local-only AE = 0.973, Federated AE = 0.972), and their ROC curves are visually almost superimposed. The ROC-AUC values were very similar for the Local-only AE (0.973) and Federated AE (0.972), and the corresponding ROC curves were visually similar. Within the synthetic-fault evaluation, this suggests that the two detectors had comparable threshold-independent discrimination. The differences observed in precision, recall, and F1-score therefore appear to be associated, at least in part, with the reconstruction-error distributions and the selected operating thresholds. This interpretation should nevertheless be regarded as preliminary because the evaluation used synthetically generated faults rather than naturally occurring labeled HVAC events.

How to cite this article:

Adejoh J, Ezeamasiohi C, Obi-Emeruwa U (October 28, 2026) Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study. *Cureus J Comput Sci* 3 : es44389-026-00344-z. DOI <https://doi.org/10.7759/s44389-026-00344-z>

Comparison with existing models

As discussed in Related Work section, the closest prior works are Wang et al. [9], who combined federated load forecasting with GMM-based anomaly prediction, and Li et al. [8], who applied FedAvg to BDG2 office buildings. A direct numerical comparison is challenging due to differing evaluation subsets and error normalization conventions. However, we provide a qualitative and contextual comparison in Table 5.

Study	Task	Dataset Scale	FL Strategy	Best Metric	Anomaly Detection
Li et al. [8]	Load Forecast	Similar office buildings	FedAvg	~10% CV-RMSE	No
Wang et al. [9]	Forecast + Anomaly	Community buildings	Adaptive FedAvg	Not reported	Residual-based GMM
Sater and Hamza [15]	Anomaly Detection	Smart building IoT	FedAvg + LSTM	Not reported	Reconstruction error
This Paper (Proof-of-Concept)	Forecast + Anomaly	5 buildings, 3 climate zones	FedAvg, FedProx, CFL, pFL	7.98% CV-RMSE (pFL)	Reconstruction AE (F1 = 0.794)

TABLE 5: Qualitative Comparison of Federated Learning Approaches for Building Load Forecasting and Anomaly Detection.

FL, Federated Learning; CV-RMSE, Coefficient of Variation of the Root Mean Square Error; LSTM, Long Short-Term Memory; AE, Autoencoder; GMM, Gaussian Mixture Model; IoT, Internet of Things.

The results are directionally consistent with prior work demonstrating potential benefits of federated learning for building energy prediction. However, direct numerical comparison is limited by differences in datasets, evaluation subsets, experimental configurations, and error-normalization procedures. Additionally, we extend their work by systematically comparing heterogeneity-aware strategies (pFL and CFL), demonstrating that these recover the accuracy lost by flat FedAvg under cross-climate divergence. Compared to Wang et al., our dedicated reconstruction-based autoencoder provides a more interpretable anomaly score than residual-based methods, though the similar ROC-AUC between federated and local detectors suggests that the primary operational benefit of federation in this task is a more balanced precision-recall trade-off.

Computational and communication efficiency

To evaluate practical deployability, we benchmarked each strategy on CPU-only hardware (Intel Xeon, 2.2 GHz, 8 GB RAM). The benchmark reflects the full federated training environment described in Hardware and Software Platforms section: 16 clients contributing gradients each round, with wall-clock time, communication rounds, and peak memory reported per client. Prediction and detection performance (Tables 2-5) remain computed only on the five evaluation buildings, consistent with the rest of the paper. Table 6 reports the total training time, communication rounds to convergence defined as the round where validation loss stabilizes within 5% of its final value, and peak memory usage per client.

How to cite this article:

Method	Training Time (min)	Communication Rounds	Peak RAM Usage (MB/client)	Total Data Transferred (MB)
Local-only (per client)	2.1	NA	450	NA
Centralized (pooled)	1.8	NA	2,100	NA
FedAvg	8.4	6	510	4.2
FedProx ($\mu = 0.01$)	8.7	6	520	4.3
Clustered FL	9.1	6	515	4.1
Personalized FL	10.2	6 + 5 fine-tuning epochs	520	4.2 + fine-tuning overhead

TABLE 6: Computational and Communication Overhead Analysis of Local, Centralized, and Federated Learning Approaches.

NA, Not Available; FL, Federated Learning.

The federated strategies converge within 6 communication rounds, taking roughly 8-10 minutes total for the 5-client simulation, approximately 4× slower than local-only training but without requiring raw data centralization. Critically, Clustered FL and Personalized FL incur only a marginal time overhead ($\leq 22\%$) compared to flat FedAvg while delivering substantial accuracy gains (e.g., a 2.0 percentage-point improvement in CV(RMSE) for pFL). The total data transferred (Table 6) remains modest, at 4.1-4.3 MB cumulatively over the 6 communication rounds to convergence (0.7 MB per round), because the models contain fewer than 2,000 parameters. These benchmarks indicate that the evaluated heterogeneity-aware FL strategies are computationally manageable in the small-scale CPU simulation used in this study. Their computational and communication requirements at full portfolio scale remain to be established, provided that the full 1,636-building portfolio is parallelized or batched into smaller federated cohorts.

Limitations and challenges

Several limitations should be considered when interpreting the findings. First, the prediction and anomaly-detection results were evaluated on five stratified buildings rather than the full BDG2 portfolio. Although 11 additional buildings participated in federated training, the reported performance metrics were calculated only for the five evaluation buildings. The study should therefore be regarded as a proof-of-concept rather than a portfolio-wide validation. The evaluated buildings span three climate zones, while the Cold climate zone was excluded because only one suitable building was available for the stratified evaluation. Larger and geographically diverse building samples are needed to determine whether the observed patterns generalize across the broader BDG2 portfolio.

Second, the anomaly-detection evaluation does not use naturally occurring labeled HVAC faults. Because BDG2 does not provide such labels, synthetic sensor drift, stuck-meter, and spike faults were injected at a fixed 12% contamination rate. The resulting metrics therefore characterize performance under the specified synthetic fault scenarios and should not be interpreted as evidence of equivalent performance on real HVAC failures. Future studies should evaluate the framework using expert-labelled fault events or datasets containing naturally occurring equipment failures.

How to cite this article:

Adejoh J, Ezeamasiohi C, Obi-Emeruwa U (October 28, 2026) Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study. Cureus J Comput Sci 3 : es44389-026-00344-z. DOI <https://doi.org/10.7759/s44389-026-00344-z>

Third, the present federated implementation does not provide a formal privacy guarantee. Raw building data remain local and only model updates are exchanged; however, differential privacy, secure aggregation, and cryptographic protections were not implemented or quantitatively evaluated. Consequently, the study demonstrates data decentralization through federated training rather than a formally quantified privacy-preserving mechanism. Future work should incorporate and evaluate appropriate privacy mechanisms, including differential privacy and secure aggregation, together with their associated accuracy and communication costs.

Fourth, no inferential statistical tests were applied to the prediction results. Because the evaluation contains only five buildings, the means and standard deviations reported in Table 2 should be interpreted descriptively. The observed differences between methods are therefore preliminary and should be validated using a larger number of independent buildings. Where sufficient samples are available, future experiments should apply paired statistical analyses across buildings or repeated evaluation splits.

Fifth, the current simulation assumes synchronous client participation and does not explicitly model stragglers, network failures, or client dropouts. Although such assumptions simplify controlled comparison of the federated strategies, they may result in more favorable convergence behavior than would occur in operational deployments.

Finally, the literature review has methodological limitations. Study screening was conducted by a single reviewer, no formal risk-of-bias or study-quality assessment was performed, and retrieval counts were not retained. The literature findings should therefore be interpreted as a structured narrative synthesis rather than a formal systematic review. Future work should use independent screening, formal quality assessment, and complete search-flow documentation.

Conclusions

This study presented a proof-of-concept evaluation of federated learning for building energy prediction and HVAC anomaly detection using the BDG2 dataset. The evaluation focused on five stratified buildings spanning three climate zones, while 11 additional buildings participated in federated training. The study compared FedAvg, FedProx, climate-zone Clustered FL, and Personalized FL with Local-only and Centralized baselines for next-hour load prediction and compared Local-only and Federated autoencoders for anomaly detection. For the prediction task, Personalized FL produced the lowest observed CV (RMSE) (7.98%) and the highest observed R^2 (0.861) among the evaluated methods. Clustered FL also produced lower observed error than flat FedAvg, whereas FedAvg and FedProx showed very similar numerical performance in this experiment. These results are consistent with the possibility that client heterogeneity influences the performance of a shared federated model and that personalization or client grouping may help address this heterogeneity. However, because only five buildings were evaluated and no inferential statistical tests were performed, these differences should be interpreted as preliminary observations rather than statistically confirmed effects.

For anomaly detection, the Federated AE achieved an F1-score of 0.794 compared with 0.748 for the Local-only AE under the synthetic fault-injection protocol. The two approaches produced very similar ROC-AUC values (0.972 and 0.973, respectively), suggesting that their threshold-independent discrimination was comparable in the evaluated setting. The observed differences in precision and recall therefore appear to be influenced by reconstruction-error distributions and threshold selection. Importantly, these findings are based on synthetically injected fault patterns and do not establish performance on naturally occurring HVAC faults. The study also evaluated computational and communication overhead in a CPU-only simulation. The federated strategies required more training time than Local-only training but converged within six communication rounds in the evaluated configuration. These measurements provide an initial indication of computational feasibility for the implemented small-scale simulation; they should not be interpreted as evidence of scalability to the full BDG2 portfolio. Overall, the findings provide an initial experimental baseline for studying federated building energy analytics under cross-building and cross-climate heterogeneity. The results support further investigation of personalized and clustered federated strategies, but larger and geographically diverse building samples are required to establish generalizability. Future research should evaluate the framework on the full or substantially larger BDG2 portfolio, include additional climate zones, use naturally occurring or expert-labeled HVAC fault events, and assess robustness under client dropout and network variability. Formal privacy mechanisms such as differential privacy and

How to cite this article:

Adejoh J, Ezeamasiohi C, Obi-Emeruwa U (October 28, 2026) Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study. *Cureus J Comput Sci* 3 : es44389-026-00344-z. DOI <https://doi.org/10.7759/s44389-026-00344-z>

secure aggregation should also be incorporated and evaluated. These extensions would provide stronger evidence regarding the accuracy, privacy, robustness, and scalability of federated building-energy analytics before practical deployment is considered.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: John Adejoh, Chibuzor Ezeamasiobi

Acquisition, analysis, or interpretation of data: John Adejoh, Ugonna Obi-Emeruwa

Drafting of the manuscript: John Adejoh

Critical review of the manuscript for important intellectual content: John Adejoh, Ugonna Obi-Emeruwa, Chibuzor Ezeamasiobi

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with [the ICMJE uniform disclosure form](#), all authors declare the following:

- **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work.
- **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work.
- **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request.

References

1. International Energy Agency (IEA): [Energy Efficiency Policy Toolkit 2025 - Buildings](#). IEA, Paris; 2026.
2. International Energy Agency (IEA): [Global Energy Review 2025 - Electricity](#). IEA, Paris; 2025.
3. REN21: [Renewables 2025 Global Status Report — Buildings](#). REN21, Paris; 2025.
4. U.S. Department of Energy: [Energy Savings Potential and RD&D Opportunities for Commercial Building HVAC Systems](#). Office of Energy Efficiency and Renewable Energy, 2017.
5. U.S. Department of Energy: [Heating, Ventilation, Air Conditioning, Refrigeration, and Water Heating](#). U.S. Department of Energy, 2020.
6. Fan C, Xiao F, Zhao Y: [A short-term building cooling load prediction method using deep learning algorithms](#). Applied Energy. 2017, 195:222-233. [10.1016/j.apenergy.2017.03.064](#)
7. Miller C, Kathirgamanathan A, Picchetti B, et al.: [The Building Data Genome Project 2: Energy meter data from the ASHRAE Great Energy Predictor III competition](#). Scientific Data. 2020, 7:368. [10.1038/s41597-020-00712-x](#)

How to cite this article:

Adejoh J, Ezeamasiobi C, Obi-Emeruwa U (October 28, 2026) Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study. Cureus J Comput Sci 3 : es44389-026-00344-z. DOI <https://doi.org/10.7759/s44389-026-00344-z>

8. Li J, Zhang C, Zhao Y, Qiu W, Chen Q, Zhang X: [Federated learning-based short-term building energy consumption prediction method for solving the data silos problem](#). Building Simulation. 2022, 15:1145-1159. [10.1007/s12273-021-0871-y](#)
9. Wang R, Yun H, Rayhana R, et al.: [An adaptive federated learning system for community building energy load forecasting and anomaly prediction](#). Energy and Buildings. 2023, 295:113215. [10.1016/j.enbuild.2023.113215](#)
10. McMahan HB, Moore E, Ramage D, Hampson S, Agüera y Arcas B: [Communication-efficient learning of deep networks from decentralized data](#). Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA. 2017. 1273-1282. [10.48550/arXiv.1602.05629](#)
11. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V: [Federated optimization in heterogeneous networks](#). Proceedings of Machine Learning Systems. 2020. 2:429-450. [10.48550/arXiv.1812.06127](#)
12. Zhao HX, Magoulès F: [A review on the prediction of building energy consumption](#). Renewable and Sustainable Energy Reviews. 2012, 16:3586-3592. [10.1016/j.rser.2012.02.049](#)
13. Wang Z, Wang Y, Zeng R, Srinivasan RS, Ahrentzen S: [Random Forest based hourly building energy prediction](#). Energy and Buildings. 2018, 171:11-25. [10.1016/j.enbuild.2018.04.008](#)
14. Sattler F, Muller KR, Samek W: [Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints](#). IEEE Transactions on Neural Networks and Learning Systems. 2021, 32:3710-3722. [10.1109/tnnls.2020.3015958](#)
15. Sater RA, Hamza AB: [A federated learning approach to anomaly detection in smart buildings](#). ACM Transactions on Internet of Things. 2021, 2:1-23. [10.1145/3467981](#)
16. Shi Y, Xu X: [Deep federated adaptation: An adaptative residential load forecasting approach with federated learning](#). Sensors. 2022, 22:3264. [10.3390/s22093264](#)
17. Fernández JD, Menci SP, Lee CM, Rieger A, Fridgen G: [Privacy-preserving federated learning for residential short-term load forecasting](#). Applied Energy. 2022, 326:119915. [10.1016/j.apenergy.2022.119915](#)
18. Zhang G, Zhu S, Bai X: [Federated learning-based multi-energy load forecasting method using CNN-attention-LSTM model](#). Sustainability. 2022, 14:12843. [10.3390/su141912843](#)

How to cite this article:

Adejoh J, Ezeamasiohi C, Obi-Emeruwa U (October 28, 2026) Federated Learning for Building Energy Prediction and Fault Detection Across Climate Zones: A Proof-of-Concept Study. Cureus J Comput Sci 3 : es44389-026-00344-z. DOI <https://doi.org/10.7759/s44389-026-00344-z>