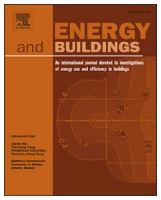
Energy & Buildings 313 (2024) 114249

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enbuild



Building consumption anomaly detection: A comparative study of two probabilistic approaches

Davor Stjelja [a,b,*], Vladimir Kuzmanovski [c], Risto Kosonen [a,d], Juha Jokisalo [a]

[a] Department of Mechanical Engineering, Aalto University, Espoo, Finland
[b] Granlund Oy, Helsinki, Finland
[c] Vaisala Oyj, Espoo, Finland
[d] College of Urban Construction, Nanjing Tech University, Nanjing, China

ARTICLE INFO

Keywords:
Anomaly detection
Drift detection
Probabilistic prediction
Conformal prediction
Building energy consumption prediction

ABSTRACT

This paper investigates the performance of two probabilistic approaches, Ensemble batch Prediction Intervals (EnbPI) a conformal prediction approach and XGBoost Location, Scale and Shape (XGBoostLSS), in predicting building energy consumption and in detecting systemic anomalies with proposed alarm matrix. The research questions focus on the effectiveness of these models in providing both point and probabilistic predictions and their utility in identifying collective anomalies. Both models showed good point and distribution prediction performance. For example, the observed point prediction had CV-RMSE in the range of 9 to 17%, outperforming recommendations from ASHRAE. Furthermore, a post-processing stage, the alarm matrix, effectively flags collective, repetitive anomalies, thus shifting focus from conventional point anomalies. The EnbPI-based method yields higher recall rates, with a trade-off of having more false alarms, while XGBoostLSS-based method excels in precision, minimizing overlooked alarms. Moreover, a robustness analysis was carried out to evaluate how these models performed when faced with training datasets containing anomalies. The robustness analysis revealed that the EnbPI-based method was more prone to overfitting, meaning its performance actually improved when the training data included some noise. On the other hand, the XGBoostLSS-based method was more stable, performing well with low levels of noise with performance drop when the noise level was high. While the findings contribute significantly to building energy consumption prediction and anomaly detection, future research could address performance in dynamic environments for the methodology and explore continual learning strategies.

1. Introduction

A well-known fact is that buildings have a significant impact on the climate [1,2]. HVAC systems alone, particularly in developed countries, are responsible for 50% of a building's energy consumption [3]. Recognizing these alarming statistics, the EU has set a goal to develop a sustainable, competitive, secure, and decarbonized energy system by 2050 [1]. One key approach to this is the digitalization of energy systems and buildings, as marked by the introduction of the EU's Smart Readiness Indicator (SRI). The SRI evaluates buildings by their capability to utilize information and communication technologies to adjust building operations to the needs of occupants and the grid, thereby improving the overall performance of buildings [1].

Technologies such as Artificial Intelligence (AI) and Machine Learning (ML) are increasingly being leveraged in building operations, enabling use cases like model predictive control, system fault detection and diagnosis, occupancy estimation and detection, and demand response, serving as integrators of different building subsystems and occupants [4,5]. Such advancements promise to enhance building performance significantly, delivering benefits ranging from improved indoor climate and lower energy consumption to better space efficiency.

According to research conducted by Katipamula and Brambley [6,7], the energy used in commercial buildings can be misspent in the range of 15 to 30% due to poorly maintained or inaccurately controlled building equipment. This study underscores the vital importance of the efficient operation and maintenance of such systems.

Moreover, seemingly small oversights in the programming of building management systems can have significant impacts on energy consumption. Studies by Abdelalim et al. and Gunay et al. [8,9] bring to

* Corresponding author.
E-mail address: davor.stjelja@granlund.fi (D. Stjelja).

https://doi.org/10.1016/j.enbuild.2024.114249
Received 15 August 2023; Received in revised form 12 April 2024; Accepted 2 May 2024
Available online 8 May 2024

Nomenclature

BDG2	Building Data Genome Project 2
BMS	Building Management System
BTE	Bayesian Target Encoding
CV-RMSE	Coefficient of variance of the root mean square error
CWC	Coverage width-based criterion
EnbPI	Ensemble batch Prediction Intervals
FMI	Finnish Meteorological Institute
GAMLSS	Generalized Additive Models for Location, Scale, and Shape
GM	Granlund Manager
LightGBM	Light Gradient-Boosting Machine
LOO	leave-one-out
LTTV	Logarithmic Transformation of Target Variable
MLE	Maximum Likelihood Estimation
NGBoost	Natural Gradient Boosting
PGBM	Probabilistic Gradient Boosting Machine
PICP	Prediction interval coverage probability
RMSE	Root mean squared error
XGBoostLSS	XGBoost Location, Scale and Shape

light that even minor control programming errors can inflate energy usage by as much as 25%.

Compounding these issues, components integral to the building's infrastructure, such as dampers, valves, and sensors, are prone to unpredictable failures [10]. Such failures can trigger excessive energy consumption, particularly if they impact various building sub-systems, thereby contributing to an overall increase in energy wastage.

Therefore, it becomes apparent that an efficient, reliable solution for data-driven anomaly detection in energy consumption data is not just beneficial but crucial. Such a solution could lead to substantial energy savings and make a significant contribution towards energy efficiency and sustainability.

Anomaly detection, also referred to as Fault Detection and Diagnosis in some literature, can be categorized into data-based and model-based methods [11,12]. Data-based methods encompass both quantitative and qualitative approaches, with a primary interest in this study being 'black box' methods that are further divided into data-mining and machine learning techniques.

Data-mining techniques often employ motif discovery and clustering. However, both have their inherent challenges [13]. Specifically, motif discovery is tailored for static data analysis, struggling in real-time data scenarios, and grappling with computational intensity and high dimensionality. This technique has been explored in various studies like [14–17], which predominantly focus on identifying infrequent patterns (discords) daily. These methods usually demand manual parameter settings. Quintana et al.'s recent work [17] enhanced a method to be parameterless, boasting a high true positive rate of 82%, but also exhibiting a high false positive rate of 65% in detecting discords.

Conversely, clustering mandates user-specified cluster numbers and isn't tailored for online data. One noteworthy application of clustering in anomaly detection can be found in [18], where the Local Outlier Factor (LOF) algorithm was used to pinpoint anomalies. Yet, clustering is often coupled with other techniques, as seen in [19,20]. For instance, Lei et al. [20] enhanced clustering with LOF, K-Nearest Neighbors (KNN), and Dynamic Time Warping (DTW) distance, formulating a dynamic method capable of detecting both point and collective anomalies. However, they noted that non-uniform events like holidays could impact their approach's anomaly detection efficacy and have not been tested on a longer dataset.

Another notable data-mining method is Association Rule Mining (ARM) [21,22], which is more effective with categorical variables and has been predominantly deployed for pinpointing anomaly causation in building subsystems with a higher number of input variables.

Machine learning techniques are further divided into classification and regression methods, depending on the problem they are trying to resolve. Both methods are supervised and necessitate labeled training datasets. However, for the task of anomaly detection, categorizing anomalous data instances is much more effortful, compared to regression task that learn and predict patterns from measured quantities, e.g., heating or electricity consumption. The regression methods minimize the discrepancy between predicted and observed values (or the reconstruction error in autoencoders), which can be utilized as the anomaly score. The key element in these approaches is determining the anomaly threshold. This process, known as thresholding, can be categorized into two main types: Static and Adaptive. The Adaptive category is further split into two subtypes: one that relies on the distribution of prediction errors, and another that focuses on predictive uncertainty.

In the paper by Yin et al. [23], novel methods for abnormal energy consumption detection are introduced, particularly suitable for fast dynamic real-time scenarios such as in industrial machines. The proposed algorithms, the "Rain Flow–based Connectivity Outlier Factor" and the "Rain Flow–based Mean Nearest Neighbor Distance Anomaly Factor" integrate time sorting and piecewise cubic Hermite interpolating polynomial methodologies, enabling them to pinpoint anomalies quickly. However, the authors have noted that these methods would have decreased performance with variable-periodic data. The thresholding method used in their approach is adaptive, based on error.

Comparatively, Munir et al. [24] proposed a deep learning-based approach for identifying point, contextual anomalies, and discords in time-series data with a static manual threshold for anomaly detection. Pan et al. [25] also presented a deep learning-based method for point anomalies detection with dynamic thresholding based on past errors. Zhang et al. [26] employed Graph Neural Networks for point anomalies detection with a manually defined static threshold.

Malhotra et al. [27] introduced a deep learning-based method for anomaly detection with adaptive thresholding based on past errors. Zhang et al. [28] utilized Xgboost and Prophet for forecasting, coupled with non-parametric dynamic thresholding for anomaly detection focusing more on recent errors.

Tambuwal et al. [29] utilized Deep Quantile Regression for anomaly detection, employing quantiles as a measure of uncertainty. In their approach, anomalies are flagged when uncertainty exceeds a predefined static threshold.

In a related vein, Beykirch et al. addressed the point anomaly detection for building heat load time series by leveraging a probabilistic forecast combination technique based on an ensemble of deterministic forecasts [30]. Their approach employed the Generalized Additive Models for Location, Scale, and Shape (GAMLSS) framework. The study yielded admirable probabilistic forecast and point anomaly detection performance. However, while GAMLSS is renowned for its adaptability and explanative nature, it comes with its set of challenges. Primarily designed for traditional regression models, GAMLSS is typically apt for small to medium-sized datasets. In situations demanding swift computations or handling of extensive data, GAMLSS might not always stand out as the optimal tool [31]. Their thresholding method is adaptive, based on uncertainty.

In the context of anomaly detection using regression methods, the research articles mentioned are summarized in a table format. The Table 1 includes three critical components for each article: the most significant information provided in that article, suggestions for enhancing the study, and the type of thresholding method employed.

Table 1
Summary of Research Articles on anomaly detection with regression methods and their thresholding method.

Reference	Most significant information from article	Drawback	Suggestions to enhance	Thresholding method
[23]	Method for real-time detection of point & collective anomalies	Proposed method performs worse with variable-periodic data like building energy consumption	Adaptability to variable-periodic data, by incorporating models that effectively handle fluctuations in energy consumption	Adaptive (error)
[24]	Deep learning-based approach for identifying point, contextual anomalies, and discords in time-series data.	Manual threshold for anomaly detection	Automatic thresholding	Static
[25]	Deep learning-based approach for point anomalies detection with dynamic thresholding	Thresholding based on past errors, can be sensitive to noise and trends	Threshold based on prediction uncertainty	Adaptive (error)
[26]	Graph Neural Networks for point anomalies detection	Manual threshold definition	Automatic thresholding	Static
[27]	Deep learning-based method for anomaly detection	Thresholding based on past errors, can be sensitive to noise and trends	Threshold based on prediction uncertainty	Adaptive (error)
[28]	Employs Xgboost and Prophet for forecasting, coupled with non-parametric dynamic thresholding for anomaly detection	Thresholding with more focus on recent errors, could be sensitive to noise and trends	Threshold based on prediction uncertainty	Adaptive (error)
[29]	Employs Deep Quantile Regression for anomaly detection, using quantiles to measure uncertainty and flagging anomalies when it surpasses a set threshold	Threshold is fixed and manual	Automatic thresholding	Static
[30]	Probabilistic forecast and anomaly detection using GAMLSS	GAMLSS is computationally inefficient and authors considered only point anomalies	Use a more efficient model and expand detection to include more than point anomalies	Adaptive (uncertainty)

Building upon the adaptive uncertainty-based thresholding approach used by Beykirch et al. [30] and aiming for computational efficiency the attention is turned to tree-based models. Tree-based models have consistently demonstrated superior performance in Kaggle competitions pertaining to building energy consumption data, such as the M5 and the GEP3 energy prediction challenge [32,33]. Reflecting this, this study employs these models to predict energy consumption and subsequently detect anomalies.

The study by Chandola et al. further expanded on the concept of anomalies, introducing the classifications of point, contextual, and collective anomalies [34]. While point anomalies indicate individual deviations, contextual and collective anomalies often reflect more profound, systemic issues that might otherwise remain undetected. In the context of this work, the focus is not on single-point anomalies, such as in [30], which represent already occurred deviations, as no actionable measures can be taken post-occurrence. The interest of this work lies in identifying systemic irregularities within buildings. By detecting these significant anomalies, systemic inefficiencies can be potentially addressed and rectified, thereby optimizing building energy consumption. In light of this, the present research introduces the alarm matrix concept to capture these more nuanced anomalies (as described further in the next section).

A review of the existing literature has aided in the development of this study. Research gaps identified from the existing literature include:

- Methods require manual setups, such as user-defined parameters or an anomaly threshold.
- Methods with adaptive threshold approaches frequently depend on the distribution of prior model errors and may be vulnerable to noise and fluctuations in trends
- Methods might not be adaptable for other tasks such as forecasting.

Informed by these identified gaps; the research questions this paper aims to tackle are crafted to offer solutions and insights to discovered gaps:

1. Probabilistic building energy consumption prediction: How effectively can probabilistic tree-based approaches perform in predicting building energy consumption and determining its expected range?
2. Detection of actionable anomalies: How can probabilistic predictions be employed for the purpose of detecting collective anomalies?

The novelty of this work centers on a unique unsupervised method designed to detect systematic anomalies in building energy consumption, which also has capabilities for energy prediction. The novelty primarily emerges from the post-processing of conditional distributions modeled using probabilistic tree-based approaches with the alarm matrix. This step enables the detection of both contextual and collective anomalies, providing a more nuanced and flexible threshold for detecting systemic irregularities in buildings. Moreover, this study stands out for its pioneering evaluation of two different approaches, namely XGBoost Location, Scale and Shape (XGBoostLSS) and Ensemble batch Prediction Intervals (EnbPI), a conformal prediction method, for generating probabilistic predictions, expanding the understanding of these techniques within this field. Additionally, the analysis included scenarios where the training datasets contained anomalies, providing insights into the comparative robustness of these models The results of this work, mark a significant step forward in the development and application of robust and efficient techniques for prediction and anomaly detection in building energy consumption data.

2. Methodology

This section introduces the devised methodology for detecting systemic anomalies in the electrical and heating energy consumption within buildings. It begins with an explanation of Bayesian target encoding, a technique used for transforming categorical features. Subsequently, the section explains two probabilistic methods for predicting energy consumption. The final part of this section is dedicated to the deployment of an alarm matrix, method for highlighting systemic anomalies.

Fig. 1 shows the workflow of this article, starting with data collection and followed by data cleaning to ensure quality inputs for subsequent stages. Feature engineering refined the predictors, and model optimization aimed to enhance model accuracy. Clean dataset analy-

Fig. 1. Sequential workflow of this study. The workflow traces the process from data collection to robustness analysis, highlighting the main stage of artificial anomaly detection assessment and essential details of each step with summarized results.

sis establishes a performance baseline for both probabilistic models, leading to the main stage of the study: testing the method to detect anomalies. A robustness analysis concludes the process, providing a metric for the consistency of model performance in cases where anomalies are present in the training dataset.

2.1. Bayesian target encoding

The majority of machine learning models are designed to work with numeric variables, meaning that categorical variables like the day of the week or the hour of the day can often lead to suboptimal outcomes. To circumvent this issue, target encoding was developed as a technique to convert these categorical variables into numerical values, thereby streamlining the modeling process. In this research, the selected method of target encoding is the Bayesian Target Encoding (BTE), known for its accuracy and efficacy as demonstrated in Kaggle competition specifically focused on energy prediction [35].

This type of target encoding, introduced in [36,37], transforms categorical variables into numerical values by calculating the mean, variance, and higher moments (like skewness) of the target variable within each category. The method involves updating a 'prior' distribution of parameters based on training data, resulting in a 'posterior' distribution, which is used to make predictions for unseen data. The statistics derived from the posterior distribution in the training dataset should be applied to the test dataset. It is important to avoid computing these statistics directly on the test dataset, as such an approach could result in data leakage.

Additionally, Bayesian target encoding allows to extract information from the intra-category distribution of the target variable. Specifically for this work, it enables the differentiation between hours during weekends or public holidays and hours on regular working days.

Given a time series dataset, $\mathcal{D} = \{(x_t, y_t)\}_{t=1}^{T}$, where x_t represents the feature vector at time t and y_t is the corresponding target variable, our preprocessing and feature engineering steps involve the following:

1. **Logarithmic Transformation of Target Variable (LTTV):**
   - The target variable y_t undergoes a logarithmic transformation to stabilize variance:

$$y'_t = \log(1 + y_t) \tag{1}$$

2. **BTE:**
   - Categorical features are grouped, and for each group G_i, statistics such as count n_i, mean $\mu_{\text{MLE},i}$, and variance $\sigma^2_{\text{MLE},i}$ are calculated using Maximum Likelihood Estimation (MLE). MLE is a statistical method that estimates the parameters of a model by maximizing the likelihood function, ensuring the estimated parameters result in a model that best explains the observed data. A Bayesian target encoding is performed, which considers both the prior information and the calculated statistics, leading to a precision-adjusted encoded feature. Using a predefined prior precision λ, the precision for each group is calculated as:

$$\text{precision}_i = \lambda + \frac{n_i}{\sigma^2_{\text{MLE},i}} \tag{2}$$

   - The encoded feature for each group G_i is then calculated as:

$$\text{BTE}(G_i) = \frac{\lambda \cdot \mu_{\text{prior},i} + \frac{n_i}{\sigma^2_{\text{MLE},i}} \cdot \mu_{\text{MLE},i}}{\text{precision}_i} \tag{3}$$

By implementing this approach, the predictive modeling capability is enhanced. The BTE, particularly, provides a nuanced understanding of time-related categories, which is crucial for energy prediction tasks. The transformation of categorical features into numerical ones through this method not only leverages the intra-category distribution but also aligns with the Bayesian framework of updating priors based on the observed data, thus leading to more accurate and robust predictions. In this work, the effect of BTE and LTTV on the performance of the models is assessed as well.

2.2. Probabilistic gradient boosting methods

Gradient boosting methods (GBM) are typically employed to yield probabilities for classification tasks. However, for regression tasks, they provide a point prediction, which can be considered as the mean of a Gaussian distribution with constant variance. For a robust probabilistic prediction, a constant variance may not suffice, which poses a challenge for GBMs. Algorithms such as Natural Gradient Boosting (NGBoost) [38], Probabilistic Gradient Boosting Machine (PGBM) [39] or XGBoostLSS [31], are trying to solve it. After evaluation, the XGBoostLSS algorithm was chosen for further exploration due to its computational efficiency and ease of implementation.

2.2.1. XGBoostLSS

An XGBoost-based probabilistic algorithm XGBoostLSS [31] is based on the original XGBoost [40] implementation, where the loss func-

tion (log-likelihood) is interpreted from a statistical perspective and the model fits both the expected value and the variance. The author borrows the approach from the well-established statistical framework of GAMLSS [41,42], where empirical risk minimization is formulated as Maximum Likelihood estimation. The algorithm uses a two-step approach, wherein the first step separate models for each distributional parameter are estimated using Maximum Likelihood, while the second step is used for updating of the already trained model. In the end, random samples are drawn from the predicted distribution, allowing prediction intervals to be derived.

2.3. Conformal prediction

Conformal prediction is a method for generating prediction intervals with a guaranteed frequentist coverage probability for any model [43,44]. Originally introduced by Vovk et al. [45], offers a distribution-free uncertainty quantification approach, with only the confidence level needing to be predefined. While conformal prediction has wide-ranging applicability across various prediction problems, the emphasis in this study is on its use in time-series prediction. Several implementations of conformal prediction for time-series prediction have been introduced in [46–49]. As an area attracting significant research interest, advancements continue to be made. Among the available techniques, the EnbPI [46,47] algorithm has demonstrated ease of implementation and notable performance, thus warranting its selection for use in this study.

2.3.1. EnbPI

EnbPI is a robust and computationally efficient algorithm for constructing prediction intervals around ensemble estimators [47]. For this approach, it is not necessary to have another dataset for calibration, like in some other conformal prediction methods.

EnbPI uses bootstrapping to fit a fixed number of estimators from subsets of training data. These estimators are central to the EnbPI methodology as they serve as the predictor models that are essential for generating initial point predictions, around which the prediction intervals are constructed. Then these estimators are used for prediction with using leave-one-out (LOO) approach, where each bootstrap estimator generates a prediction for each data point in the training set, but in a way that each prediction is made by a model that was not trained on that particular data point. This process results in LOO predictors and LOO residuals.

In the prediction phase, EnbPI aggregates these LOO predictors for each test data point to compute the central prediction. It then uses the past LOO residuals to build prediction intervals around this central prediction. These intervals offer a range of plausible values for the predicted response variable, thus providing an indication of the uncertainty around the prediction.

2.4. Alarm matrix

As outlined in the introduction, this study's emphasis is directed toward contextual and collective anomalies rather than point anomalies. Thus, not every point that falls outside the predicted interval is automatically classified as an anomaly. To address such instances more accurately, a post-processing algorithm, known as the alarm matrix, has been implemented. The alarm matrix examines each point that deviates from the predicted range. This examination involves looking at similar time intervals from the previous two days and from the corresponding weekdays two weeks prior (Fig. 2). Here, similar hours refer to the hour immediately before and after the time when the observed consumption is detected to fall outside the predicted range. If there were similar point anomalies in recent history, they would be marked as anomalous. The alarm matrix is suitable for the online implementation of this work's approach.

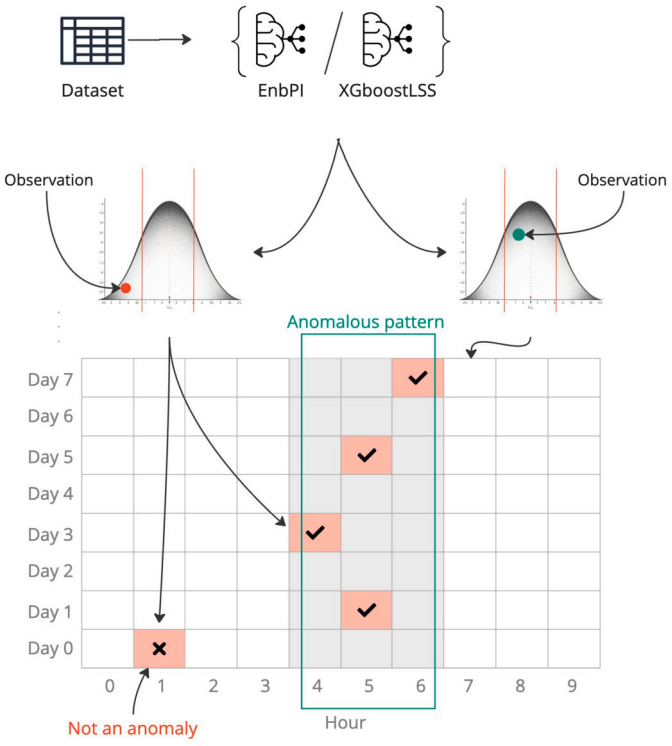


Fig. 2. Conceptual design of the alarm matrix approach that takes the probabilistic output and condition it to past time points. If anomalous pattern appears then all detected anomalies are flagged as predicted anomalies, while the rest discarded.

3. Results

The results section showcases a comparative analysis of the proposed method for systemic anomaly detection, which includes an alarm matrix and a probabilistic prediction model. It begins with a description of the dataset used for the initial experiments. This is followed by an outline of the evaluation metrics employed for both point and interval predictions, as well as the classification metrics used for anomaly detection. The process of preprocessing and feature engineering of the dataset is explained next, along with the optimization of parameters and configuration of the model. The probabilistic prediction models, XGBoostLSS and EnbPI with Light Gradient-Boosting Machine (LightGBM), are then examined using datasets of clean electrical and heating consumption. The effects of target encoding and logarithmic transformation on model performance are also evaluated. Subsequently, the paper compares both prediction models using the proposed alarm matrix method for detecting systemic anomalies against a reference method [28]. Lastly, the robustness of both probabilistic models with the alarm matrix is analyzed.

3.1. Experimentation dataset

The dataset used in this paper comprises hourly records of district heating and electricity energy consumption for a variety of public and commercial buildings located in Finland, as well as a selection of buildings from the Building Data Genome Project 2 (BDG2) [50], an open dataset of energy meter data. The Finnish data was gathered over a span of four years, between 2018 and 2022, from Granlund Manager (GM) the facility management system [51]. For the purpose of this research, periods of stable energy consumption without significant drift were selected from the Finnish dataset. Such periods, spanning 12 months for training followed by an additional 2 months for testing, were selected to ensure data consistency and avoid significant fluctuations in consumption patterns. Additionally, the BDG2 dataset, which includes data from

Table 2
Characteristics of experimentation dataset.

Consumption type	Building	Area (m^2)	Volume (m^3)	Correlation coeff.	Consumption (kWh)			Temperature (°C)		
					Min	Avg	Max	Min	Med	Max
Heating	School 1	3300	13100	-0.83	9	190	550	-15.7	4.4	32.8
	School 2	7500	34000	-0.91	0	260	940	-25.9	5.1	32.8
	School 3	5900	27300	-0.83	0	135	470	-15.7	4.4	32.8
	Senior home	6500	27700	-0.86	0	156	590	-25.9	5.4	31.2
	Hog office Marlena (scaled)	12240	n/a	-0.63	0	0.26	0.87	-28.9	8.9	35
	Hog office Bessie (scaled)	8976	n/a	-0.91	0	0.25	1.0	-28.9	8.9	35
Electricity	University building	11700	48740	0.02	21	158	388	-26	7.1	31.4
	Shopping mall 1	18900	83970	0.25	17	85	172	-26.4	4.7	30.6
	Shopping mall 2	16200	102500	0.13	29	146	287	-11.1	6.3	30.3
	Senior home	6500	27700	0.05	0	119	258	-26.4	7.6	31.8
	Hog education Bruno (scaled)	13972	n/a	0.04	0.09	0.53	0.92	-28.9	10.6	35
	Hog office Corrie (scaled)	11103	n/a	0.05	0.02	0.34	0.90	-28.9	10.6	35

2016 and 2017 from buildings in North America and Europe, has been incorporated to enhance the diversity of the data and to facilitate reproducibility of the work, as the Finnish buildings data is not available for public use.

To evaluate the performance of consumption anomaly detection, an additional version of this dataset was created, in which artificial anomalies were introduced during the test period, which is explained in section 3.6. These anomalies take on various forms to simulate different conditions and challenges the models may encounter, enabling a rigorous test of their ability to detect these anomalies.

The dataset is divided into two subsets: heating and electricity consumption, gathered from various buildings. Table 2 details the characteristics of these subsets. The heating dataset includes Finnish buildings like schools and a senior home, alongside office buildings from BDG2. The electricity dataset comprises data from Finnish locations such as shopping malls, a senior home, and a university building, supplemented by an educational and an office building from BDG2. This varied collection, encompassing educational, commercial, residential, and office spaces from Finnish and BDG2 sources, offers a comprehensive view of energy consumption patterns across different building types.

Table 2 also presents minimum, average, and maximum consumption alongside temperature data. Buildings in both Finland and BDG2 are in cold climates, with BDG2 buildings experiencing comparatively warmer summers. Pearson's correlation coefficient analysis has been done between temperature and consumption. This analysis shows a strong negative correlation between heating consumption and temperature, while electricity consumption exhibits a weaker correlation, slightly higher in shopping malls due to cooling and/or auxiliary heating demands. BDG2 buildings' consumption data are scaled, omitting unit-based absolute consumption figures.

Temperature data, a key factor affecting energy usage, varies based on the building's location. For Finnish buildings, temperature readings were sourced from the nearest Finnish Meteorological Institute (FMI) weather station. Conversely, the BDG2 dataset includes its temperature data, providing an accurate environmental context for energy consumption analysis and serving as a primary input in model inference.

This combination of a clean dataset and one with artificially introduced anomalies ensures a comprehensive evaluation of the studied methods, encompassing their performance under normal conditions as well as their capacity to detect anomalies.

3.2. Evaluation metrics

In the evaluation of model performance on a clean dataset, the focus is on a mix of point prediction and probabilistic prediction metrics. For point predictions, Root mean squared error (RMSE) and Coefficient of variance of the root mean square error (CV-RMSE) are utilized. While RMSE serves as a widely-accepted metric for regression models due to its ability to measure average prediction error magnitudes, CV-RMSE is particularly valuable for energy consumption data. Given the potential for wide variations in energy consumption magnitudes, CV-RMSE offers a normalized measure, ensuring consistent comparability across diverse datasets and scenarios.

Given this work's emphasis on probabilistic predictions, it is vital to evaluate the accuracy of predicted intervals. Predicted intervals are evaluated using Prediction interval coverage probability (PICP) and Coverage width-based criterion (CWC) methods, since these metrics have been widely utilized in the field of energy predictions already [52,53,49].

When it comes to assessing the effectiveness of the proposed anomaly detection methodology, classification metrics are adopted. These include precision, recall, and the F1 score. Collectively, these metrics provide a comprehensive evaluation of the model's performance in detecting anomalies.

For reference, these metrics are defined as follows:

- **RMSE**: This is a widely used measure of the differences between the values predicted by a model and the values actually observed. It's calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

  where y_i are the observed values, $\hat{y}_i$ are the predicted values, and n is the number of observations.
- **CV-RMSE**: This is used to assess the model's prediction accuracy, and it's calculated as:

$$CVRMSE = \frac{RMSE}{\bar{y}} \times 100 \tag{5}$$

  where $\bar{y}$ is the mean of the observed values.
- **PICP**: This is a measure of a model's reliability in providing a certain prediction interval. It is the percentage of times that the actual value falls within the predicted interval.

$$PICP = \frac{1}{n}\sum_{i=1}^{n} c_i \tag{6}$$

  where c_i is an indicator function, defined as:

$$c_i = \begin{cases} 1, & y_i \in [L_i, U_i] \\ 0, & y_i \notin [L_i, U_i] \end{cases} \tag{7}$$

  Here, y_i is the actual value, and $[L_i, U_i]$ is the predicted interval for each observation i.
- **CWC**: This is a performance measure that balances the PICP and the width of the prediction interval. It is defined as:

$$CWC = (1 - PINAW) \times e^{-\eta \times (PICP - (1-\alpha))^2} \tag{8}$$

where $PINAW$ is the Prediction Interval Normalized Average Width, calculated as:

$$PINAW = \frac{1}{nR}\sum_{i=1}^{n}(U_i - L_i), \qquad R = y_{\max} - y_{\min} \tag{9}$$

Also in the CWC equation, α is the confidence level and η is a scaling factor that magnifies the differences between $PICP$ and α. The value for η is set empirically and in this case, the value of 0.3 is used as in [49].

- **Precision**: Precision is a measure of the model's accuracy in terms of its ability to return only relevant instances. It's calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

where TP is the number of true positives and FP is the number of false positives.

- **Recall**: Recall, also known as sensitivity or true positive rate, measures the model's ability to identify all relevant instances. It's calculated as:

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

where TP is the number of true positives and FN is the number of false negatives.

- **F1 Score**: The F1 score is the harmonic mean of precision and recall, offering a balance between the two metrics. It's particularly useful in situations where the data have an uneven class distribution. The F1 score is calculated as:

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

It ranges between 0 (worst) and 1 (best).

3.3. Data pre-processing and feature engineering

In the data pre-processing phase, consumption values were merged with temperature data for each respective hour. Outliers, potentially resulting from metering issues, were identified and removed based on a threshold set at six standard deviations from the median. The missing values resulting from this process were then imputed using a rolling median from the previous week.

Once the dataset was cleaned, the next step was feature engineering. Lagged values and a moving average of temperature were derived for a range of one to 72 hours prior to each prediction point. In addition, datetime features, such as the day of the week, hour, month, and public holidays, were also included. To leverage categorical features effectively, and to capture their mutual interactions, Bayesian target encoding was employed.

3.4. Parameter optimization and model configuration

For this research, the XGBoostLSS python package (version 0.2) was used [54]. Developed using PyTorch [55], this version offers the advantages of automatic differentiation for gradients and Hessians. This facilitates accelerated computation and allows for the introduction of advanced distributions. Moreover, it introduces a feature that evaluates the optimal distribution for specific datasets—a critical prerequisite before leveraging this method [56].

Effectively applying XGBoostLSS required determining an appropriate distribution. To achieve this, the integrated feature for distribution selection was used, which employs a negative log likelihood (NLL) approach. After evaluating the majority of buildings, it became evident that the Beta distribution was the most suitable choice, and it was consequently adopted.

In the scope of this research, EnbPI was implemented utilizing the MAPIE Python package (version 0.6.4) [57] as a wrapper. Within this implementation, LightGBM served as the predictor [58]. Notably, LightGBM has empirically showcased swift performance, with outcomes that align closely with those of XGBoost. As detailed in Section 2.3.1, the necessity of an underlying prediction model for EnbPI is thoroughly explained, underscoring its pivotal role in the algorithm's overall functioning.

Table 3
Hyperparameter ranges for different algorithms.

Algorithm	Hyperparameter	Range
LightGBM (EnbPI)	max_depth	2-30
	num_leaves	200-5000
	n_estimators	100-3000
	min_data_in_leaf	100-1000
	learning_rate	0-1
	reg_alpha	0-1
	reg_lambda	0-1
	subsample	0-1
	colsample_bytree	0-1
XGBoostLSS	param_dict	1e-5 - 1
	max_depth	1-10
	gamma	1e-8 - 40
	subsample	0.2-1.0
	colsample_bytree	0.2-1.0
	min_child_weight	1e-8 - 500

The choice of LightGBM as the predictor model was based on its proven computational efficiency and its ability to deliver robust results, especially in scenarios demanding high-speed data processing without compromising predictive accuracy. This aligns well with the requirements of EnbPI, which necessitates a fast yet accurate predictor to efficiently handle multiple bootstrapping and LOO iterations. LightGBM's architecture, known for its scalability and lower memory usage, particularly enhances the performance of EnbPI in handling large datasets and complex time series patterns. Furthermore, its empirical performance, evidenced through comparative studies with similar algorithms like XGBoost, validated its suitability for this research, ensuring that the prediction intervals generated by EnbPI are both reliable and computationally feasible.

Integral to the EnbPI implementation was the determination of critical parameters, namely the bootstrap length and the number of resamplings. In determining the optimal parameters for bootstrap length and number of resamplings, an empirical iterative experimentation approach was utilized. This involved systematically adjusting these parameters and assessing model performance based on accuracy and computational efficiency. Consequently, the bootstrap length was set at 3500 for electricity data and 750 for heating data, aligning with the unique temporal patterns of each dataset. The number of resamplings was established at 150, balancing computational feasibility and precision of prediction intervals. This method ensured optimal configuration for the model, addressing specific data characteristics and maintaining robust performance. It is noteworthy that both XGBoostLSS and EnbPI's underlying LightGBM predictor were carefully optimized using a representative dataset. This process allowed us to ascertain the most effective hyperparameters for each model. The specific hyperparameters and their explored ranges for both models are detailed in Table 3.

The code is available for reproducibility on GitHub: https://github.com/sirdawar/ProbabilisticBuildingAnomaly.

3.5. Comparative analysis on clean dataset

This section presents a comparative analysis of the two models, XGBoostLSS and EnbPI, focusing on their performance in predicting energy consumption. Both models are applied to the clean dataset, which represents normal, undisturbed conditions of energy usage in various types of buildings.

The goal of this comparative analysis is twofold. First, it aims to evaluate the accuracy of each model in producing point predictions of

Table 4
Comparative analysis of two algorithms on clean heating and electricity dataset.

Consumption type	Building	Algorithm	RMSE	CV-RMSE (%)	PICP	CWC
Heating	School 1	EnbPI	0.047	9.56	0.82	0.50
		XGBoostLSS	0.049	9.88	0.98	0.66
	School 2	EnbPI	0.035	10.03	0.92	0.80
		XGBoostLSS	0.038	11.13	0.99	0.57
	School 3	EnbPI	0.047	11.78	0.85	0.60
		XGBoostLSS	0.049	12.48	1.00	0.44
	Senior home	EnbPI	0.033	10.04	0.85	0.66
		XGBoostLSS	0.035	10.56	0.97	0.75
	Hog office Marlena (BDG2)	EnbPI	0.043	10.43	1.00	0.57
		XGBoostLSS	0.043	10.58	0.99	0.69
	Hog office Bessie (BDG2)	EnbPI	0.060	13.07	0.90	0.76
		XGBoostLSS	0.060	13.06	0.98	0.71
Electricity	University building	EnbPI	0.050	14.41	0.86	0.62
		XGBoostLSS	0.050	14.41	0.91	0.74
	Shopping mall 1	EnbPI	0.054	12.76	0.88	0.71
		XGBoostLSS	0.052	12.25	0.94	0.76
	Shopping mall 2	EnbPI	0.052	10.97	0.95	0.74
		XGBoostLSS	0.055	11.49	0.94	0.66
	Senior home	EnbPI	0.057	16.86	0.87	0.67
		XGBoostLSS	0.058	16.92	0.98	0.71
	Hog education Bruno (BDG2)	EnbPI	0.070	15.78	0.88	0.74
		XGBoostLSS	0.063	14.23	0.98	0.77
	Hog office Corie (BDG2)	EnbPI	0.072	17.39	0.88	0.68
		XGBoostLSS	0.067	16.13	0.94	0.79

Fig. 3. Comparative Visualization of Heating Consumption Predictions and Prediction Intervals by XGBoostLSS and EnbPI Models in School 1 (Top) and Senior home (Bottom). (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

energy consumption, as measured by the RMSE and CV-RMSE. Second, it assesses the effectiveness of the models in providing reliable probabilistic predictions, as evaluated by the PICP and the CWC.

These metrics provide a comprehensive assessment of the performance of each model, capturing both their accuracy in predicting energy consumption and their reliability in generating prediction intervals. By comparing these metrics across the two models, this analysis will provide insights into their relative strengths and limitations, offering guidance for the selection of the most suitable model for predicting energy consumption under normal conditions.

A detailed comparative analysis of the performance metrics across different buildings and models is depicted in Table 4 and in Figs. 3 and 4. Figures serve as illustrative representations of the performance of EnbPI (red) and XGBoostLSS (blue) models in predicting heating and electricity consumption, respectively, across different buildings. They provide a visual depiction of the interplay between point predictions and prediction intervals, allowing for a better understanding of model performance.

Analyzing the point prediction performance of the models reveals a closely matched accuracy for energy consumption. For heating pre-

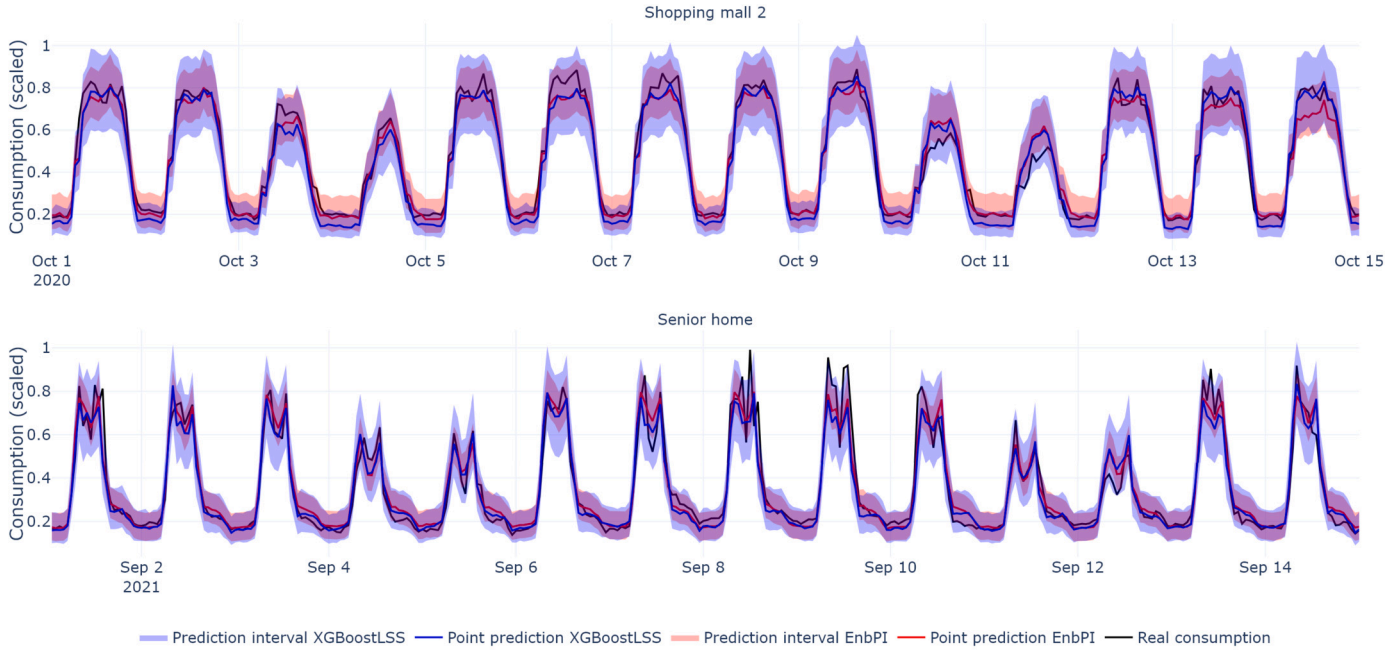


Fig. 4. Comparative Visualization of Electricity Consumption Predictions and Prediction Intervals by XGBoostLSS and EnbPI Models in Shopping mall 1 (Top) and Senior home (Bottom).

dictions, the EnbPI model slightly surpasses the XGBoostLSS model, while the better model for electricity prediction fluctuates across different buildings. However, a more comprehensive measure of predictive quality is provided by the CV-RMSE metric. According to ASHRAE guidelines [59] a CV-RMSE of less than 25% signifies a good congruence of building simulation to actuality. In this analysis, the heating CV-RMSE ranges from 9% to 13%, while electricity ranges from 11% to 17%, suggesting both models yield high-quality predictions. The patterns exhibited in the figures are in alignment with the results outlined in the tabulated data. For example, in the graphical illustrations, much like in the table, point predictions generated by both the EnbPI and XGBoostLSS models closely mirror the actual energy consumption, reinforcing the precision of these models.

When evaluating the performance of prediction intervals for heating consumption, the EnbPI model exhibits a more conservative outcome, with the PICP ranging between 0.82 and 1.00 (0.86-0.95 for electricity). On the other hand, XGBoostLSS achieves higher values, in the range of 0.97 to 1 (0.91-0.98). It is important to note that while the desired confidence level is set at 0.95, neither of the models perfectly achieves this target, except EnbPI in one case. Nonetheless, XGBoostLSS approaches this benchmark slightly more closely in the majority of cases. Upon considering the CWC, which penalizes both under and overcoverage, the results appear to be divided. Evaluation prediction intervals from Figs. 3 and 4 confirm conclusions that the XGBoostLSS model generates wider prediction intervals, capturing most of the data points, thus ensuring greater coverage. While, the EnbPI model, with its more conservative, narrower prediction intervals, occasionally misses some outliers. In summary, each model shows instances of superior and inferior performance, with the outcomes largely depending on the specific building.

Upon a review of the comparative analysis, it is apparent that predicting heating consumption might be slightly more accurate than predicting electrical consumption. This is presumably due to the more direct correlation between outdoor temperature and heating consumption, especially in the context of North European climates. On the other hand, while the prediction of electrical consumption is marginally less precise, the results are still of high quality. The lower correlation of outdoor temperature with electrical consumption introduces a layer of complexity, but does not compromise the overall efficacy of predictive models.

This observed trend regarding prediction accuracy and interval width is consistent across all buildings, as demonstrated in Figs. 3 and 4, which represent heating and electricity consumption, respectively. In the context of anomaly detection, EnbPI's narrower prediction intervals could potentially give rise to more false alarms, whereas the broader intervals from XGBoostLSS could potentially overlook some actual anomalies due to their wider coverage. Therefore, the selection of the better model depends on the application, striking a balance between minimizing false alarms and mitigating the risk of overlooking real anomalies. Consequently, a thorough analysis of the anomaly detection capabilities of the proposed approach, comparing these two methods, is carried out in section 3.6 to explore these aspects in greater detail.

3.5.1. Bayesian Target Encoding (BTE) and Logarithmic Transformation of Target Variable (LTTV) effect on the model performance

This work has further investigated the effects of BTE and LTTV on the model performance, as shown in Table 5. Regarding heating consumption, point prediction accuracy has either slightly increased or remained at the same level when not using BTE or LTTV. However, distribution prediction with XGBoostLSS can significantly decline, as indicated by negative CWC values. In the realm of electrical consumption prediction, where time-related (categorical) features play a pivotal role, the inclusion of BTE has been particularly beneficial in enhancing model performance by increasing the impact and interactions of these features. While heating consumption is more closely related to continuous variables like temperature, where benefits of using BTE and LTTV are less pronounced, if not negligible. This analysis has shown that the EnbPI model demonstrates robustness against variations in scale, in contrast to the XGBoostLSS model, where using LTTV is crucial for stabilizing variance.

3.6. Comparative analysis with artificial anomalies

In this subsection, the first step involves outlining the process for incorporating artificial anomalies into the dataset and the characterization of these anomalies. Following this, the focus then turns to a detailed examination of the results stemming from this comparative analysis.

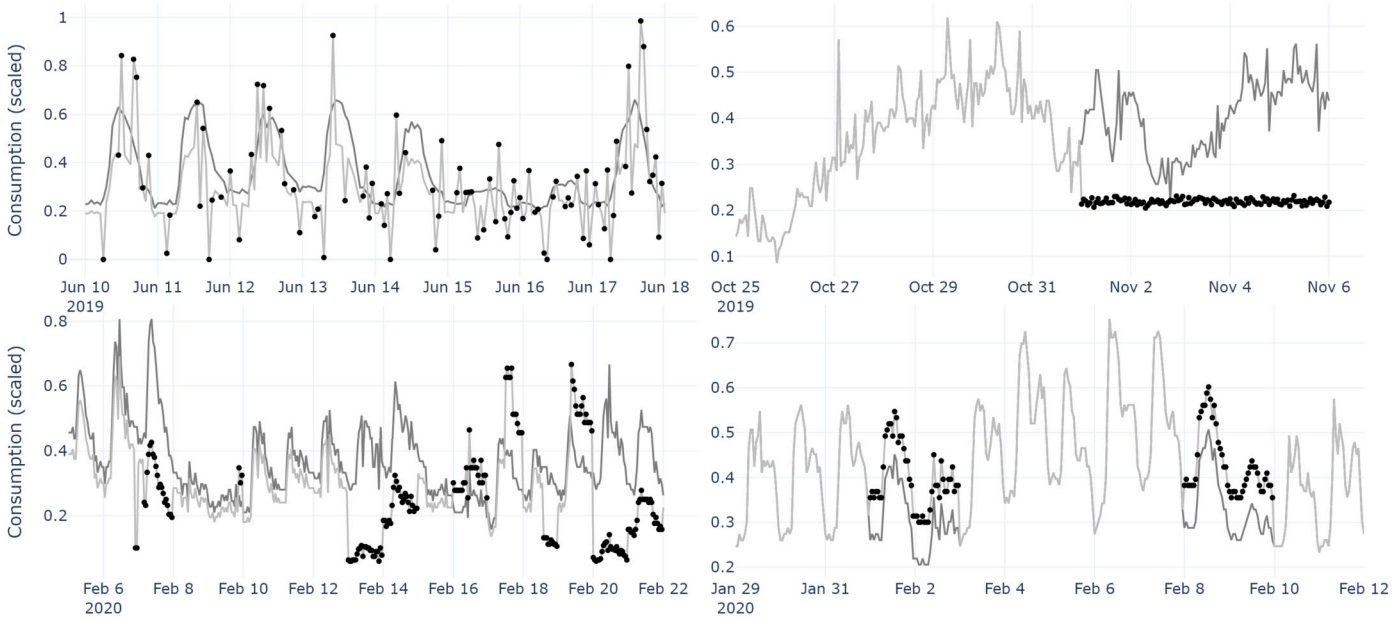


Fig. 5. Illustrations of artificially introduced anomalies in energy consumption. Clockwise from upper left: a random anomaly depicting a potential meter reading error; a sudden halt of major building operations; an alteration in the weekend profile; and a significant change in the building's operational pattern. Clean consumption is represented in gray, altered consumption in silver, and altered hours are marked with black dots.

Table 5
Comparative analysis of models on clean heating and electricity dataset with or without BTE and LTTV.

Consumption type	Building	Model	BTE with LTTV		No BTE with LTTV		No BTE no LTTV	
			CV-RMSE (%)	CWC	CV-RMSE (%)	CWC	CV-RMSE (%)	CWC
Heating	Hog office Marlena	EnbPI	10.43	0.57	9.43	0.59	9.14	0.65
		XGBoostLSS	10.58	0.69	12.90	0.62	16.40	-19.87
	Hog office Bessie	EnbPI	13.07	0.76	11.62	0.80	11.50	0.80
		XGBoostLSS	13.06	0.71	13.12	0.71	15.24	-1.18
Electricity	Hog education Bruno	EnbPI	15.78	0.74	13.60	0.67	13.83	0.69
		XGBoostLSS	14.23	0.77	14.80	0.77	59.30	-2.02
	Hog office Corie	EnbPI	17.39	0.68	21.61	0.65	21.50	0.62
		XGBoostLSS	16.13	0.79	19.06	0.75	1131.44	-0.01

In order to assess the effectiveness of anomaly detection, artificial anomalies are added to the clean dataset. The anomalies introduced into the dataset are designed to mimic potential real-world issues with building energy consumption. These issues can stem from various factors such as changes in Building Management System (BMS) schedules and setpoints, manual overrides, significant user behavior changes within the buildings, or issues related to smart meter and data transfer.

Fig. 5 provides visual illustrations of the anomalies introduced, showcasing the diverse range of challenges encountered in real-world building energy consumption scenarios.

The evaluation of the anomaly detection using the alarm matrix methodology and the probabilistic models utilizes the classification metrics specified in section 3.2. Although the data, anomalies, and their detection are represented on an hourly basis, this study concentrates on identifying collective anomalies at the daily level. Hence, an instance is considered a true positive only when an anomaly and its detection arise within the same day. Any divergences from this scenario are accounted for appropriately.

This study utilizes a non-parametric dynamic thresholding method, an adaptive error-based approach for anomaly detection, as outlined in the 2021 work by Zhang et al. ([28]), as a reference methodology. This method was integrated with the same LightGBM algorithm that was previously applied in the EnbPI analysis. Throughout this text, this combined methodology will be referred to as the "reference" method.

The comparative analysis, as detailed in Table 6, evaluates the performance of a proposed method, incorporating an alarm matrix with an underlying model that can be either EnbPI or XGBoostLSS, against a reference method. The focus of this analysis is on the detection of artificially induced anomalies. Findings indicate that the proposed method, regardless of whether it employs EnbPI or XGBoostLSS as its underlying model, demonstrates a slightly higher efficacy in anomaly detection within the heating dataset compared to the electricity dataset. This finding indicates a potential disparity in the effectiveness of these methods across different types of datasets. However, the results for the electricity dataset should not be overlooked as they still show a considerable level of effectiveness, particularly with respect to recall.

Precision, indicative of daily anomaly detection, tends to be higher when utilizing the proposed method with the XGBoostLSS model as its basis. While recall, representing the proportion of actual anomalies that the model can identify, is slightly better when the EnbPI model underlies the proposed method. Even for electricity consumption, the recall values are notably high, suggesting that a significant majority of artificial anomalies are correctly identified.

The differences in precision across various models and cases, which range from 0.28 to 0.85 for EnbPI and from 0.33 to 1.00 for XGBoostLSS-based method, indicate that there could be a significant number of false positives in certain situations. The lower precision range of the EnbPI-based method, may be attributed to its more conservative nature as noted in the clean dataset analysis. While this conservatism can lead to a lower rate of false negatives (improved recall), it also implies a higher possibility of false positives, thereby affecting precision. Conversely, XGBoostLSS-based method, with its wider

Table 6
Comparative evaluation of a proposed method, employing either EnbPI or XGBoostLSS models, against a reference method in detecting artificial anomalies within heating and electricity datasets.

Consumption type	Building	Model	Precision	Recall	F1 Score
Heating	School 1	EnbPI	0.30	1.00	0.47
		XGBoostLSS	0.62	0.83	0.71
		Reference	0.91	0.61	0.73
	School 2	EnbPI	0.71	1.00	0.83
		XGBoostLSS	1.00	1.00	1.00
		Reference	0.72	0.34	0.47
	School 3	EnbPI	0.35	1.00	0.51
		XGBoostLSS	0.59	1.00	0.75
		Reference	0.59	0.37	0.45
	Senior home	EnbPI	0.69	1.00	0.81
		XGBoostLSS	0.97	0.97	0.97
		Reference	0.73	0.46	0.56
	Hog office Marlena (BDG2)	EnbPI	0.85	0.49	0.63
		XGBoostLSS	1.00	0.33	0.50
		Reference	0.90	0.75	0.82
	Hog office Bessie (BDG2)	EnbPI	0.44	0.88	0.58
		XGBoostLSS	0.61	0.88	0.72
		Reference	0.84	0.62	0.72
Electricity	University building	EnbPI	0.72	1.00	0.84
		XGBoostLSS	1.00	1.00	1.00
		Reference	0.86	0.61	0.71
	Shopping mall 1	EnbPI	0.28	1.00	0.44
		XGBoostLSS	0.33	1.00	0.50
		Reference	0.83	0.69	0.75
	Shopping mall 2	**EnbPI**	0.59	0.86	0.70
		XGBoostLSS	0.48	0.91	0.62
		Reference	0.82	0.41	0.55
	Senior home	EnbPI	0.29	1.00	0.45
		XGBoostLSS	0.68	1.00	0.81
		Reference	1.00	0.82	0.90
	Hog education Bruno (BDG2)	EnbPI	0.29	1.00	0.45
		XGBoostLSS	0.83	0.71	0.77
		Reference	0.71	0.86	0.78
	Hog office Corie (BDG2)	EnbPI	0.52	0.92	0.67
		XGBoostLSS	0.62	0.96	0.75
		Reference	0.76	0.50	0.60

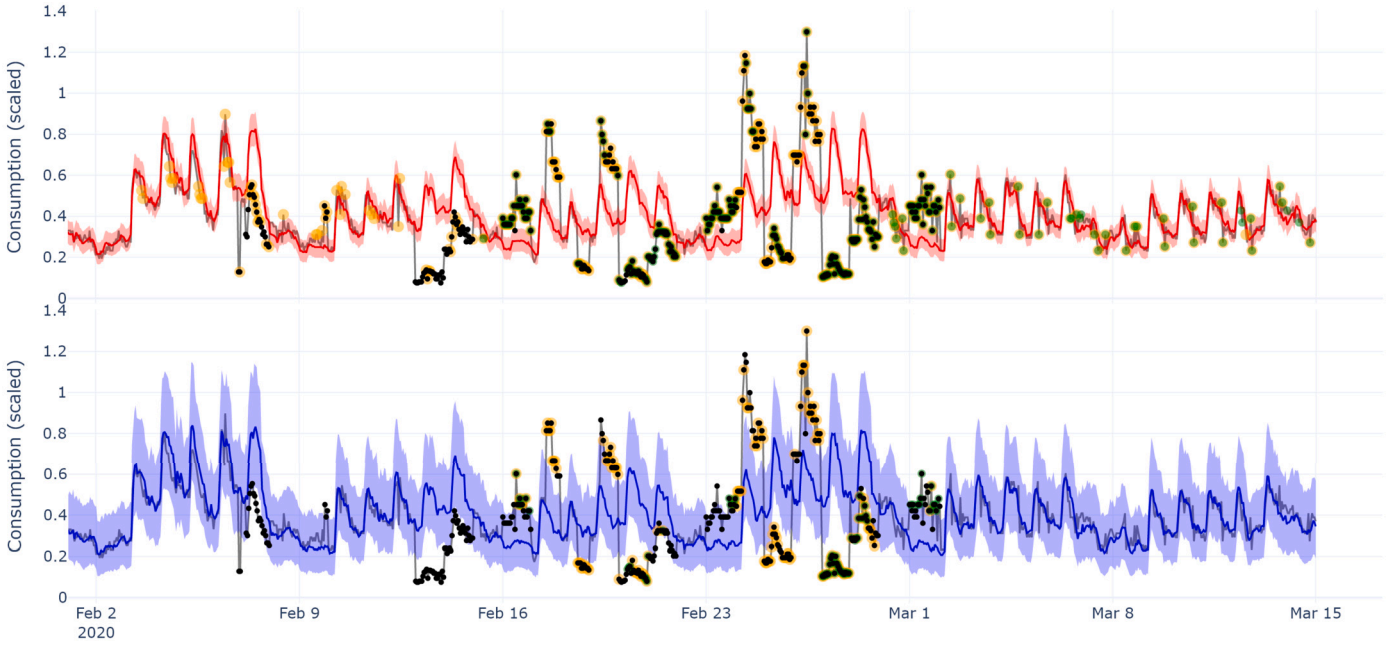


Fig. 6. Comparative visualization of detected anomalies in heating consumption using the alarm matrix method with underlying EnbPI (red) and XGBoostLSS (blue) models. This representation is based on data from School Building 3. Black dots indicate artificially introduced anomalies. Recurrent anomalies are denoted by yellow dots for anomalies occurring at least two consecutive days and green dots for anomalies persisting on at least two consecutive instances of the same weekday.

prediction intervals, generally delivers a better balance of precision and recall, reflected in its superior F1 score across most scenarios.

In Table 6, the method with the highest overall score for each case is highlighted in bold. Predominantly, the XGBoostLSS-based method

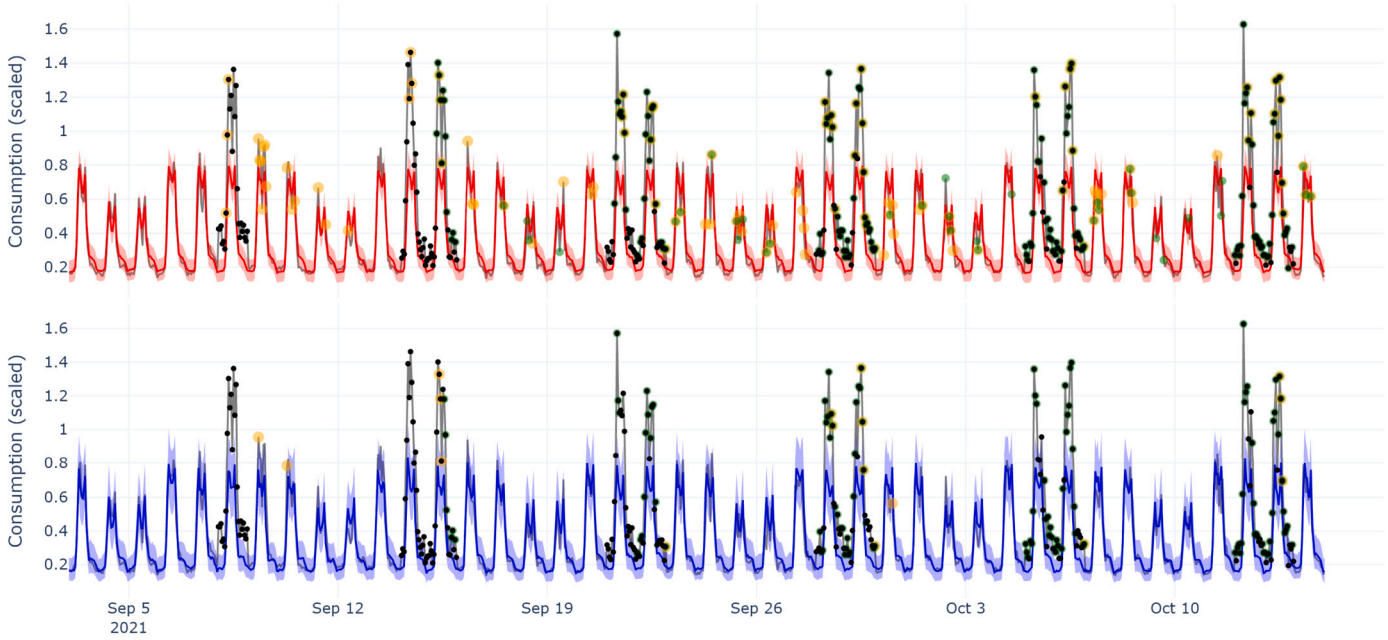


Fig. 7. Comparative visualization of detected anomalies in electricity consumption using the alarm matrix method with underlying EnbPI (red) and XGBoostLSS (blue) models. This representation is based on data from Senior home. Black dots indicate artificially introduced anomalies. Recurrent anomalies are denoted by yellow dots for anomalies occurring at least two consecutive days and green dots for anomalies persisting on at least two consecutive instances of the same weekday.

outperforms the others, followed by the reference method, with the EnbPI-based method trailing. A crucial distinction to note is that the reference method adapts its threshold based on new data, whereas the probabilistic algorithms of EnbPI and XGBoostLSS are applied in a static manner in this study, not updating their prediction range. The continual learning of these methods is not within the scope of this research. When comparing EnbPI- and XGBoostLSS-based methods to the reference method, it is observed that the reference method often exhibits high precision but low recall. This indicates that the LightGBM combined with the thresholding method from Zhang et al. ([28]) tends to be conservative in anomaly detection. While it accurately identifies anomalies when they are detected, it also misses some genuine anomalies due to this conservative approach.

Therefore, while all methods demonstrate capabilities in anomaly detection, XGBoostLSS-based method might offer a slight edge for this purpose. However, the ultimate method selection would largely depend on the specific parameters and objectives of the implementation scenario.

Referring to the visual depiction provided in Figs. 6 and 7, it becomes clear that the conclusions drawn from the table are reflected similarly. It can be seen that the EnbPI-based method, marked in red, may have a slight edge in terms of recall, successfully identifying a higher proportion of actual anomalies. However, this comes at the cost of a higher rate of false positives, which is observable from the additional red dots. Conversely, the XGBoostLSS-based method, represented by blue dots, demonstrates a better balance between precision and recall, aligning with the results observed in Table 6.

3.7. Comparative analysis of robustness

In order to assess the robustness of methods on varying signal-to-noise ratios, artificial noise is added to the clean dataset. The analysis is designed to incorporate artificial noise according to a well established practices for investigating models' robustness under data integrity attacks [60,61]. The consumption values are altered on a randomly selected time segments (event length) by a noise sampled from a normal distribution with parameters μ – mean (noise level) and σ – standard variation that represent the centering and scale of the distribution.

The analysis of the robustness of methods is performed over all selected buildings by imputation of an artificial noise that is categorized as *low*, *medium* and *high*. The categories differ from each other by the level of the noise added to the noiseless time series and the number of events/anomalies each applied to three consecutive hours (Table 7). The noise added to the base time series is sampled from a normal distribution centered (mean) around the underlying time series values and with a scale (σ) equal to the noise level. The noisy events are sampled uniformly across the time range of the training dataset (a year) and the decision is made to reflect coverage of approx. 2%, 4% and 15% of the hours (or approx. 20%, 30% and 60% of the number of days if account the uniformity of the sampling) for the three noise categories, respectively. The test dataset for each building is kept as defined in the previous subsection and the results are reported for that dataset.

The robustness is analyzed through two perspectives: potential overfit on a clean dataset and performance retention over different levels of noise. The results summarized through *F1 score* performance metric, as shown in Fig. 8, depict variability of models' behavior across different buildings and consumption types (electricity and heating). However, the EnbPI-based method clearly gains in performance as the noise increases in the training datasets, with couple of exceptions in the high level of noise. That means that the method is prone to overfit for the task of anomaly detection, whereby the noise in the training dataset serves as a regularization during the learning process. As opposite, the XGBoostLSS-based method shows less overfitting to the clean training dataset, while retains the performance to similar levels as without noise, except for couple of cases where the training datasets were degraded with high level of noise. Finally, the XGBoostLSS-based method shows better performance independently of the noise presence and levels. More detailed results are given in Appendix A, Figs. A.9 and A.10 where along the *F1 score*, the *Precision* and *Recall* are provided, as well.

4. Discussion

The findings of this study are essential in the broader context of research objectives. A significant aspect of the investigation involves answering the pivotal question: 'How effectively can probabilistic tree-based approaches perform in predicting building energy consumption

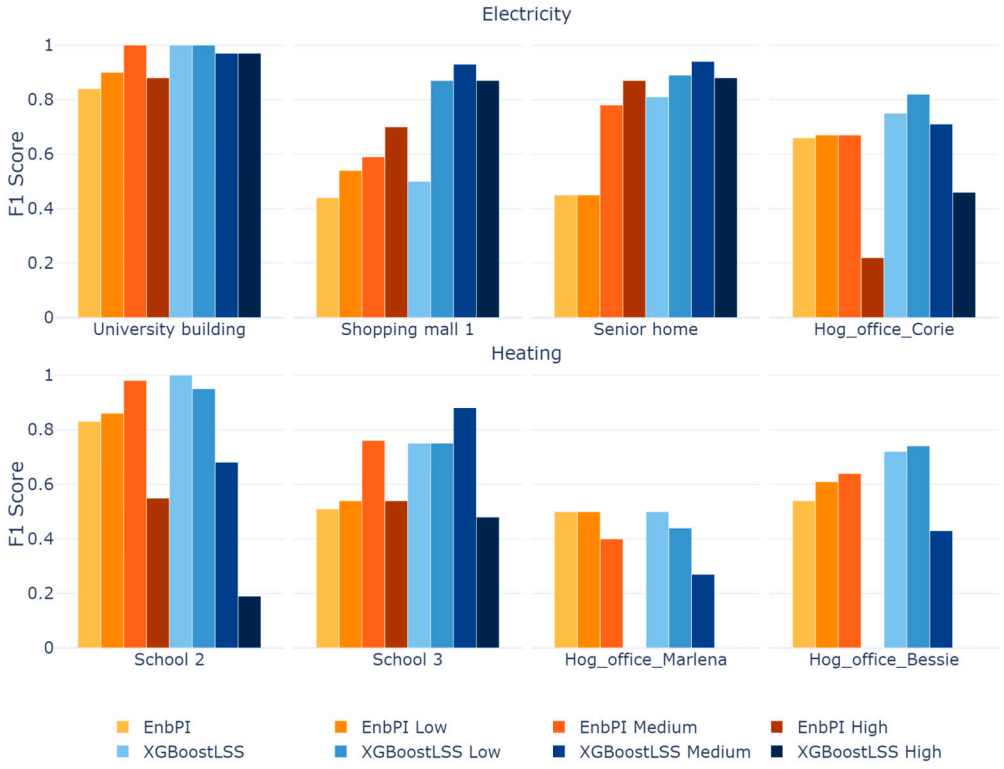


Fig. 8. Comparative Analysis of F1 Score Performance: EnbPI- vs. XGBoostLSS-based methods Across Various Noise Levels in Training Datasets (No Noise, Low, Medium, High).

Table 7
Summary and characterization of noise categories applied on the training datasets. **Noise level** is a scale of a random value sampled from a normal distribution, **# events** is the number of events over the time course of train datasets (uniformly sampled) and **event length** is the length of each event in hours.

Category	Noise level	# events	Event length
Low	0.1	72	3
Medium	0.5	110	3
High	0.75	250	3

and determining its expected range? To this end, the study contrasts two tree-based probabilistic approaches EnbPI and XGBoostLSS for predicting building energy consumption and formulates a mechanism for detecting significant anomalies. Outdoor temperature and date-time variables were incorporated into the model to predict heating and electrical energy consumption with commendable accuracy. The CV-RMSE was observed to be within the range of approximately 9 to 13% for heating and 11-17% for electricity, results which outperformed the ASHRAE's recommendations [59]. The predicted distribution for both methods was found to be realistic, although neither model consistently achieved the desired confidence level of 0.95, but close to it. The XGBoostLSS model predicted broader intervals, while EnbPI was slightly more conservative.

Model performance of methods is confirmed by the robustness analysis, showing that EnbPI-based method tends to overfit on clean data and therefore act over-confidently on unseen data. Namely, the EnbPI-based method signals much more false anomalies (reduced precision) on unseen data when training on a clean dataset. However, as the noise increases, the confidence is corrected and the false alarms get filtered out, leading to better precision and a bit lower recall - share of the true anomalies detected by the model (Appendix A). XGBoostLSS-based method, on the other side, shows less affected performances as the noise in the training data increases, except on buildings that exhibit highly stationary time series Appendix A. Conclusively, the latter is not sensitive to the presence of noise and can capture a significant amount of anomalies present in unseen data, which aligns to the findings of the study presented in [61]. Therefore, one should be very carefully revising the training dataset when using EnbPI-based, unlike use of the XGBoostLSS-based method. Overall, both approaches showed good capabilities in point and interval prediction affirmatively addressing the first research question concerning the predictive performance of tree-based probabilistic approaches in energy consumption modeling. This is valuable as it confirms that these models can generate dependable predictions with uncertainty associated with it, which can subsequently be applied to anomaly detection or forecasts.

In addition, this work tested the effects of Bayesian Target Encoding and Logarithmic Transformation of Target Variable on these models' performance. We found that BTE is particularly useful for electricity consumption prediction, which relies heavily on date-time features, but it is less impactful for heating consumption prediction. Moreover, the EnbPI model proved more robust against variations in scale compared to XGBoostLSS, for which the inclusion of LTTV is significant in terms of variance stabilization."

The second research question explored the potential for probabilistic predictions to assist in the identification of collective anomalies. An additional post-processing stage termed the alarm matrix was instituted post-prediction to eliminate point anomalies and flag collective, repetitive, and meaningful anomalies. Essentially, this process scrutinized neighboring hours from the past two days and two analogous weekdays for any consumption measurement that deviated from the expected range, subsequently marking such instances as anomalies.

The proficiency of the alarm matrix in conjunction with probabilistic predictions was evaluated by artificially introducing collective anomalies, modeled on realistic building cases. The method was optimized for an online scenario where the previous day's energy consumption is assessed, with the aim of accurately marking anomalous days. To verify its efficiency, precision, recall, and the F1 score were evaluated on a

day-to-day basis. The EnbPI-based method demonstrated a higher recall rate, indicating it is effective at identifying most of the relevant cases, but this may come with a trade-off of having more false alarms. Whereas XGBoostLSS-based method excelled in precision, resulting in fewer overlooked alarms. Both methods were capable of detecting recurrent anomalies, the choice between the two hinging on the use-case severity and method fine-tuning.

In both prediction and anomaly detection, heating proved to be a more straightforward case, yielding superior results, while electricity showed moderately lower outcomes. The correlation between district heating energy consumption and outdoor temperature, coupled with the predictable behavior of heating systems, accounted for this discrepancy. In contrast, electricity consumption, not only influenced by equipment schedules but also occupant behavior of a more stochastic nature, exhibited a more random pattern.

In the comparative analysis involving artificial anomalies, the study compared the proposed method incorporating an alarm matrix and underlying models (XGBoostLSS and EnbPI) against a reference method. This reference approach adopted error-based adaptive thresholding as outlined in Zhang et al. ([28]), and employed the LightGBM with the same hyperparameters as used in the EnbPI approach. The reference model demonstrated good results. A significant aspect to note is its ability to continuously update its threshold with new data from the test dataset, providing it with a certain advantage over the main methods of this study. It is important to clarify that while the XGBoostLSS- and EnbPI-based methods in this research did not include a similar threshold updating mechanism, there are alternative methods for updating. These alternative updating mechanisms were not explored within the scope of this study, but their potential incorporation could suggest that XGBoostLSS-based and EnbPI-based methods might achieve improved performance under different updating conditions.

The implementation of the XGBoostLSS model was more straightforward, requiring only hyperparameter optimization after distribution selection, a capability readily available in the Python package. For the implementation of EnbPI, the selection of a primary predictive algorithm is a requisite. LightGBM was the chosen algorithm due to its remarkable speed and performance characteristics. Once this selection was completed, hyperparameter optimization was conducted in a manner analogous to the process used for XGBoostLSS. Subsequent to this, the bootstrap length was determined, followed by the establishment of the number of resamplings necessary for performing conformal prediction.

The findings of this study not only contribute to the existing body of knowledge on building energy consumption prediction and anomaly detection, but they also offer significant advancements to the larger discourse within this field. Historically, the dominant approach has revolved around point prediction methods and assumptions of Gaussian distributions when dealing with distributions. Additionally, anomaly thresholds were typically established either statically or adaptively using previous prediction errors, with considerably less emphasis on incorporating uncertainty into the thresholding process. This study, however, moves beyond these traditional paradigms. It emphasizes the significant role of collective and systemic anomalies, a facet often overlooked in favor of detecting point anomalies. The need to focus on these collective anomalies is particularly critical in building environments, thus underscoring the practical implications of this work.

While this study has laid a significant foundation, it also reveals areas that need further exploration. Future research should concentrate on the following key areas:

- The adaptation of the method to the dynamic environments of buildings, where usage patterns, control systems, and climate conditions are constantly changing. Investigating approaches such as continual learning or ensemble techniques that combine models trained on both old and new data will be crucial. This research should aim to enhance model flexibility and accuracy in the face of environmental variability.
- Managing anomalies in the training dataset is critical. Research should focus on whether these anomalies should be excluded, modified, or incorporated into the training process. This involves developing strategies to identify and appropriately respond to such data inconsistencies, ensuring they do not compromise model performance.
- Another promising direction is the exploration of a broader range of probabilistic methods for building energy predictions. Specifically, the potential of Bayesian methods warrants investigation as they may offer advantages over the current frequentist approaches. This exploration should aim to assess their efficacy in different predictive scenarios and their ability to handle uncertainties more effectively.
- Lastly, there is a need to develop more generalizable models that can efficiently manage challenges like dynamic environments, missing data for new buildings, or anomalies in the training dataset. Future research should focus on creating and validating models that are trained on extensive datasets from a broad spectrum of buildings. The goal is to achieve a model that is both robust and adaptable to a wide range of building types and conditions.

By addressing these areas, future research can significantly enhance the field of building energy prediction and anomaly detection, leading to models that are more adaptable, accurate, and broadly applicable.

In order to make this work reproducible code was made available on GitHub: https://github.com/sirdawar/ProbabilisticBuildingAnomaly.

5. Conclusion

This work culminates in answers to the central research questions that were initially posed. The first question explored whether proposed probabilistic approaches, specifically Ensemble batch Prediction Intervals (EnbPI) a conformal prediction method using Light Gradient-Boosting Machine (LightGBM) and XGBoost Location, Scale and Shape (XGBoostLSS), could deliver strong performance in both point prediction and the probabilistic prediction of building energy consumption. The CV-RMSE for these models was between 9 to 13% for heating and 11 to 17% for electricity, which significantly outperformed the minimum point prediction standards recommended by ASHRAE. Furthermore, they succeeded in accurately characterizing the distribution of predictions through their generated intervals. These results affirmatively respond to the initial research question.

The second research question probed the application of probabilistic prediction for the detection of collective anomalies. The alarm matrix, a novel post-processing stage proposed in this study, proved effective in highlighting collective, repetitive, and meaningful anomalies from the outputs of probabilistic models. This innovative shift away from the conventional focus on point anomalies towards collective ones represents a significant contribution to building energy management.

In comparing the methods based on EnbPI and XGBoostLSS, each exhibited its unique strengths. However, overall, XGBoostLSS-based method demonstrated superior performance. The EnbPI-based method exhibited a higher recall rate, effectively identifying a greater proportion of relevant cases, though this could lead to an increase in false alarms. Additionally, it has proven to be resilient to variations in data magnitude, maintaining consistent performance across different scales of data. Conversely, XGBoostLSS-based method showcased superior precision, thereby reducing the likelihood of overlooked alarms in anomaly detection. It's important to note that this level of precision was achieved through the log transformation of the target variable (LTTV), a key step for stabilizing variance in the method. The choice between these two approaches will largely depend on the severity of the specific use-case and the extent of method fine-tuning needed.

Additionally, a robustness analysis was carried out to evaluate how these methods performed when faced with training datasets containing anomalies. This analysis revealed that the EnbPI-based method tended to overfit, while the XGBoostLSS-based method demonstrated greater stability in these conditions.

Despite these promising findings, the study acknowledges certain limitations. Unexplored areas include the performance of the methodology in dynamic building environments, strategies for continual learning, and the creation of a more generalized model applicable to a wide range of buildings. Future research could delve into these aspects and investigate alternative probabilistic approaches, such as Bayesian methods.

In conclusion, this study has provided noteworthy contributions to the field of building energy consumption prediction and anomaly detection. It has provided encouraging answers to the initial research questions, and it has opened up promising avenues for further exploration in this field.

CRediT authorship contribution statement

Davor Stjelja: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Vladimir Kuzmanovski:** Writing – review & editing, Validation, Methodology, Investigation, Conceptualization. **Risto Kosonen:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Juha Jokisalo:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Davor Stjelja reports financial support was provided by Granlund Oy. Davor Stjelja reports a relationship with Granlund that includes: employment.

Code availability

In order to make this work reproducible, the code was made available on GitHub: https://github.com/sirdawar/ProbabilisticBuildingAnomaly.

Data availability

Code and public data is shared on Github available in manuscript. Some of the data used is confidential and not public.

Acknowledgements

The authors would like to express their sincere gratitude for the support and resources provided for this research. This work was part of the ITEA4 project 20219 IML4E and was generously funded by Business Finland.

Appendix A

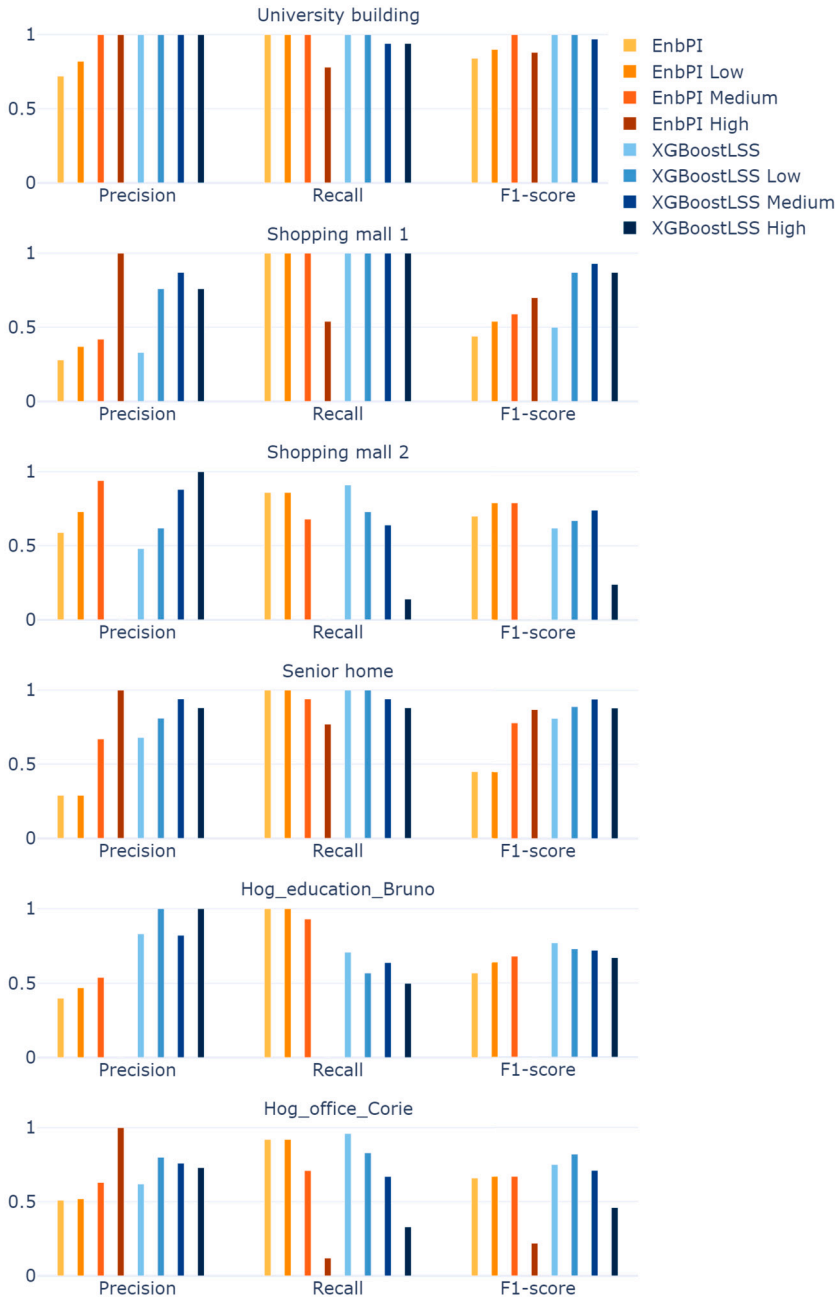


Fig. A.9. Comparative Analysis of Precision, Recall and F1-Score Performance: EnbPI vs. XGBoostLSS Models Across Various Noise Levels in Electricity Training Datasets (No Noise, Low, Medium, High)).

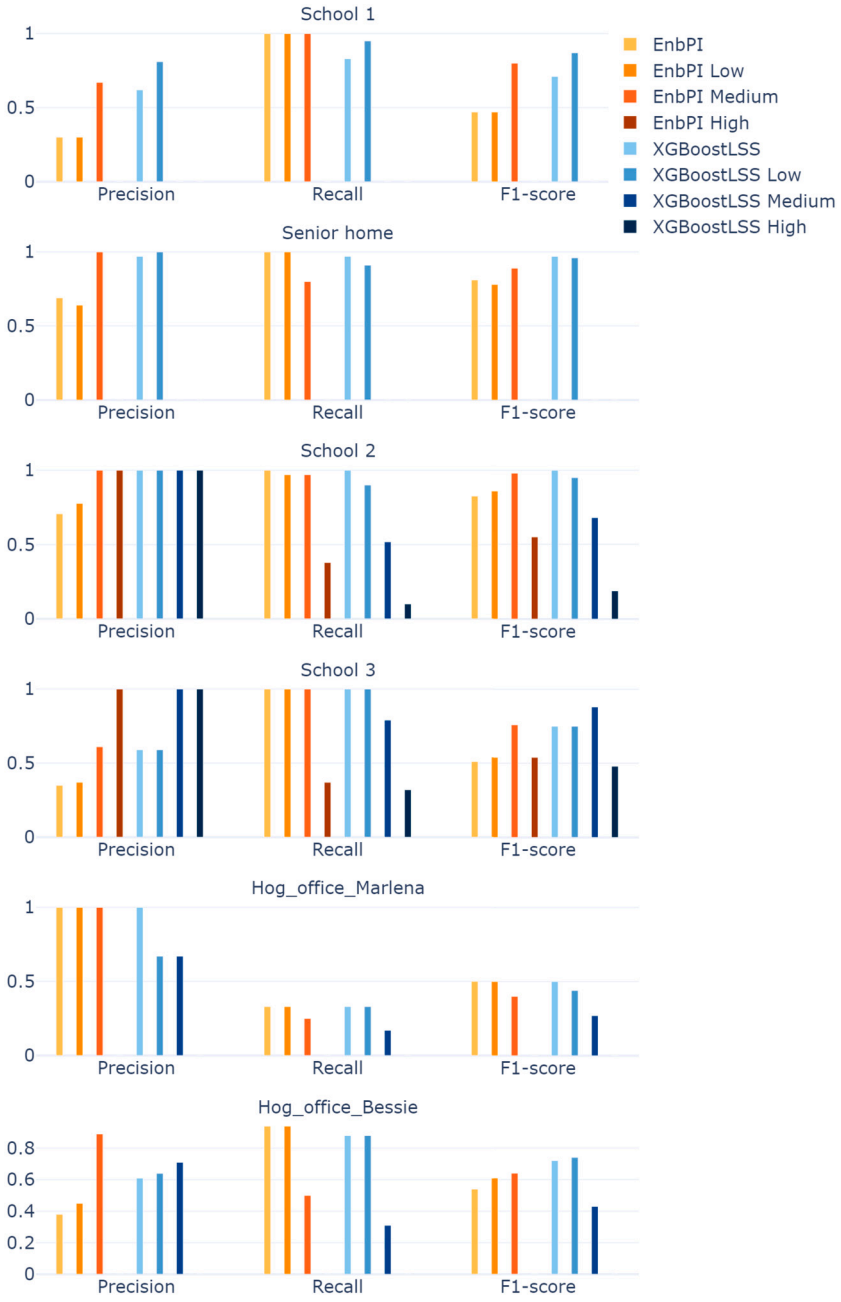


Fig. A.10. Comparative Analysis of Precision, Recall and F1-Score Performance: EnbPI vs. XGBoostLSS Models Across Various Noise Levels in Heating Training Datasets (No Noise, Low, Medium, High).

References

[1] European Union, Directive 2018/844 of the European Parliament and of the Council of 30 May 2018 amending Directive 2010/31/EU on the energy performance of buildings and Directive 2012/27/EU on energy efficiency, Off. J. Eur. Union (2018), L 156/75 – L 156/91.

[2] U.N. Environment, International Energy Agency, Towards a zero-emission, efficient, and resilient buildings and construction sector, Tech. Rep., UN Environment and International Energy Agency, 2017.

[3] L. Pérez-Lombard, J. Ortiz, C. Pout, A review on buildings energy consumption information, Energy Build. 40 (3) (2008) 394–398, https://doi.org/10.1016/j.enbuild.2007.03.007.

[4] J.-H. Choi, B. Yang, C.W. Yu, Artificial intelligence as an agent to transform research paradigms in building science and technology, Indoor Built Environ. (2021) 1420326X2110176, https://doi.org/10.1177/1420326X211017694.

[5] Z. Wang, J. Liu, Y. Zhang, H. Yuan, R. Zhang, R.S. Srinivasan, Practical issues in implementing machine-learning models for building energy efficiency: moving beyond obstacles, Renew. Sustain. Energy Rev. 143 (February 2021) 110929, https://doi.org/10.1016/j.rser.2021.110929.

[6] S. Katipamula, M. Brambley, Review article: methods for fault detection, diagnostics, and prognostics for building systems—a review, Part I, HVAC & R Res. 11 (1) (2005) 3–25, https://doi.org/10.1080/10789669.2005.10391123.

[7] S. Katipamula, M. Brambley, Review article: methods for fault detection, diagnostics, and prognostics for building systems—a review, Part II, HVAC & R Res. 11 (2) (2005) 169–187, https://doi.org/10.1080/10789669.2005.10391133.

[8] A. Abdelalim, W. O'Brien, Z. Shi, Development of Sankey diagrams to visualize real HVAC performance, Energy Build. 149 (2017) 282–297, https://doi.org/10.1016/j.enbuild.2017.05.040.

[9] B. Gunay, W. Shen, C. Yang, Characterization of a building's operation using automation data: a review and case study, Build. Environ. 118 (2017) 196–210, https://doi.org/10.1016/j.buildenv.2017.03.035.

[10] W. Kim, S. Katipamula, A review of fault detection and diagnostics methods for building systems, Sci. Technol. Built Environ. 24 (1) (2018) 3–21, https://doi.org/10.1080/23744731.2017.1318008.

[11] M.S. Mirnaghi, F. Haghighat, Fault detection and diagnosis of large-scale HVAC systems in buildings using data-driven methods: a comprehensive review, Energy Build. 229 (2020) 110492, https://doi.org/10.1016/j.enbuild.2020.110492.

[12] S.P. Melgaard, K.H. Andersen, A. Marszal-Pomianowska, R.L. Jensen, P.K. Heiselberg, Fault detection and diagnosis encyclopedia for building systems: a systematic review, Energies 15 (12) (2022) 4366, https://doi.org/10.3390/en15124366.
[13] C. Fan, F. Xiao, Z. Li, J. Wang, Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: a review, Energy Build. 159 (2018) 296–308, https://doi.org/10.1016/j.enbuild.2017.11.008.
[14] C. Miller, Z. Nagy, A. Schlueter, Automated daily pattern filtering of measured building performance data, Autom. Constr. 49 (2015) 1–17, https://doi.org/10.1016/j.autcon.2014.09.004.
[15] J.Y. Park, E. Wilson, A. Parker, Z. Nagy, The good, the bad, and the ugly: data-driven load profile discord identification in a large building portfolio, Energy Build. 215 (2020) 109892, https://doi.org/10.1016/j.enbuild.2020.109892.
[16] A. Capozzoli, M.S. Piscitelli, S. Brandi, D. Grassi, G. Chicco, Automated load pattern learning and anomaly detection for enhancing energy management in smart buildings, Energy 157 (2018) 336–352, https://doi.org/10.1016/j.energy.2018.05.127.
[17] M. Quintana, T. Stoeckmann, J.Y. Park, M. Turowski, V. Hagenmeyer, C. Miller, ALDI++ automatic and parameter-less discord and outlier detection for building energy load profiles, Energy Build. 265 (2022) 112096, https://doi.org/10.1016/j.enbuild.2022.112096.
[18] H. Rashid, P. Singh, Monitor an abnormality detection approach in buildings energy consumption, in: 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC), 2018, pp. 16–25.
[19] P. Xue, Z. Zhou, X. Fang, X. Chen, L. Liu, Y. Liu, J. Liu, Fault detection and operation optimization in district heating substations based on data mining techniques, Appl. Energy 205 (2017) 926–940, https://doi.org/10.1016/j.apenergy.2017.08.035.
[20] L. Lei, B. Wu, X. Fang, L. Chen, H. Wu, W. Liu, A dynamic anomaly detection method of building energy consumption based on data mining technology, Energy 263 (2023) 125575, https://doi.org/10.1016/j.energy.2022.125575.
[21] Z.J. Yu, F. Haghighat, B.C.M. Fung, L. Zhou, A novel methodology for knowledge discovery through mining associations between building operational data, Energy Build. 47 (2012) 430–440, https://doi.org/10.1016/j.enbuild.2011.12.018.
[22] D.F. Motta Cabrera, H. Zareipour, Data association mining for identifying lighting energy waste patterns in educational institutes, Energy Build. 62 (2013) 210–216, https://doi.org/10.1016/j.enbuild.2013.02.049.
[23] S. Yin, H. Yang, K. Xu, C. Zhu, S. Zhang, G. Liu, Dynamic real–time abnormal energy consumption detection and energy efficiency optimization analysis considering uncertainty, Appl. Energy 307 (2022) 118314, https://doi.org/10.1016/j.apenergy.2021.118314.
[24] M. Munir, S.A. Siddiqui, A. Dengel, S. Ahmed, DeepAnT: a deep learning approach for unsupervised anomaly detection in time series, IEEE Access 7 (2019) 1991–2005, https://doi.org/10.1109/ACCESS.2018.2886457.
[25] H. Pan, Z. Yin, X. Jiang, High-dimensional energy consumption anomaly detection: a deep learning-based method for detecting anomalies, Energies 15 (17) (2022) 6139, https://doi.org/10.3390/en15176139.
[26] Z. Zhang, Y. Chen, H. Wang, Q. Fu, J. Chen, Y. Lu, Anomaly detection method for building energy consumption in multivariate time series based on graph attention mechanism, PLoS ONE 18 (6) (2023) e0286770, https://doi.org/10.1371/journal.pone.0286770.
[27] P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, G. Shroff, LSTM-based encoder-decoder for multi-sensor anomaly detection, arXiv:1607.00148, Jul. 2016.
[28] W. Zhang, L. Wang, X. Zhao, Y. Liu, RobustProphet: time series anomaly detection method, in: 2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), IEEE, Fuzhou, China, 2021, pp. 157–161.
[29] A.I. Tambuwal, D. Neagu, Deep quantile regression for unsupervised anomaly detection in time-series, SN Comput. Sci. 2 (6) (2021) 475, https://doi.org/10.1007/s42979-021-00866-4.
[30] M. Beykirch, T. Janke, I. Tayeche, F. Steinke, in: Probabilistic Forecast Combination for Anomaly Detection in Building Heat Load Time Series, Nov. 2021, arXiv:2107.10828.
[31] A. März, XGBoostLSS – an extension of XGBoost to probabilistic forecasting, arXiv:1907.03178, Aug. 2019 [cs, stat].
[32] C. Miller, P. Arjunan, A. Kathirgamanathan, C. Fu, J. Roth, J.Y. Park, C. Balbach, K. Gowri, Z. Nagy, A.D. Fontanini, J. Haberl, The ASHRAE great energy predictor III competition: overview and results, Sci. Technol. Built Environ. 26 (10) (2020) 1427–1447, https://doi.org/10.1080/23744731.2020.1795514.
[33] C. Miller, B. Picchetti, C. Fu, J. Pantelic, Limitations of machine learning for building energy prediction: ASHRAE Great Energy Predictor III Kaggle competition error analysis, Sci. Technol. Built Environ. 28 (5) (2022) 610–627, https://doi.org/10.1080/23744731.2022.2067466.
[34] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: a survey, ACM Comput. Surv. 41 (3) (2009) 15:1–15:58, https://doi.org/10.1145/1541880.1541882.
[35] ASHRAE Great Energy Predictor III https://kaggle.com/competitions/ashrae-energy-prediction, 2019.
[36] A. Slakey, D. Salas, Y. Schamroth, Encoding categorical variables with conjugate Bayesian models for WeWork lead scoring engine, arXiv:1904.13001, Apr. 2019.
[37] M. Larionov, Sampling techniques in Bayesian target encoding, arXiv:2006.01317, Nov. 2020.
[38] T. Duan, A. Avati, D.Y. Ding, K.K. Thai, S. Basu, A.Y. Ng, A. Schuler, NGBoost: natural gradient boosting for probabilistic prediction, arXiv:1910.03225, Jun. 2020 [cs, stat].
[39] O. Sprangers, S. Schelter, M. de Rijke, Probabilistic gradient boosting machines for large-scale probabilistic regression, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 1510–1520, arXiv:2106.01682, https://doi.org/10.1145/3447548.3467278.
[40] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[41] R.A. Rigby, D.M. Stasinopoulos, Generalized additive models for location, scale and shape, J. R. Stat. Soc., Ser. C, Appl. Stat. 54 (3) (2005) 507–554, https://doi.org/10.1111/j.1467-9876.2005.00510.x.
[42] N. Klein, T. Kneib, S. Lang, Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data, J. Am. Stat. Assoc. 110 (509) (2015) 405–419, https://doi.org/10.1080/01621459.2014.912955.
[43] K.P. Murphy, Probabilistic Machine Learning: Advanced Topics, Adaptive Computation and Machine Learning Series, The MIT Press, Cambridge, Massachusetts, 2023.
[44] A.N. Angelopoulos, S. Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification, arXiv:2107.07511, Dec. 2022.
[45] V. Vovk, A. Gammerman, C. Saunders, Machine-learning applications of algorithmic randomness, in: Proceedings of the Sixteenth International Conference on Machine Learning, 1999, pp. 444–453.
[46] C. Xu, Y. Xie, Conformal prediction interval for dynamic time-series, in: International Conference on Machine Learning, PMLR, 2021, pp. 11559–11569.
[47] C. Xu, Y. Xie, Conformal prediction for time series, IEEE Trans. Pattern Anal. Mach. Intell. (2023) 1–22, https://doi.org/10.1109/TPAMI.2023.3272339.
[48] M. Zaffran, A. Dieuleveut, O. Féron, Y. Goude, J. Josse, Adaptive Conformal Predictions for Time Series, Feb. 2022, arXiv:2202.07282.
[49] V. Jensen, F.M. Bianchi, S.N. Anfinsen, Ensemble conformalized quantile regression for probabilistic time series forecasting, IEEE Trans. Neural Netw. Learn. Syst. (2022) 1–12, https://doi.org/10.1109/TNNLS.2022.3217694, arXiv:2202.08756.
[50] C. Miller, A. Kathirgamanathan, B. Picchetti, P. Arjunan, J.Y. Park, Z. Nagy, P. Raftery, B.W. Hobson, Z. Shi, F. Meggers, The building data genome project 2, energy meter data from the ASHRAE Great Energy Predictor III competition, Sci. Data 7 (1) (2020) 368, https://doi.org/10.1038/s41597-020-00712-x.
[51] Granlund Manager, Granlund Oy, Jul. 2023.
[52] H. Quan, D. Srinivasan, A. Khosravi, Uncertainty handling using neural network-based prediction intervals for electrical load forecasting, Energy 73 (2014) 916–925, https://doi.org/10.1016/j.energy.2014.06.104.
[53] Y. Shen, X. Wang, J. Chen, Wind power forecasting using multi-objective evolutionary algorithms for wavelet neural network-optimized prediction intervals, Appl. Sci. 8 (2) (2018) 185, https://doi.org/10.3390/app8020185.
[54] A. März, XGBoostLSS - an extension of XGBoost to probabilistic forecasting, https://github.com/StatMixedML/XGBoostLSS, Jul. 2023.
[55] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: an imperative style, high-performance deep learning library, arXiv:1912.01703, https://doi.org/10.48550/arXiv.1912.01703, Dec. 2019.
[56] A. März, T. Kneib, Distributional gradient boosting machines, arXiv:2204.00778, https://doi.org/10.48550/arXiv.2204.00778, Apr. 2022.
[57] V. Taquet, V. Blot, T. Morzadec, L. Lacombe, N. Brunel, MAPIE: an open-source library for distribution-free uncertainty quantification, arXiv:2207.12274, https://doi.org/10.48550/arXiv.2207.12274, Jul. 2022.
[58] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017.
[59] R. American, Society of Heating, G. Air Conditioning Engineers (Atlanta, Ashrae Guideline 14-2014: Measurement of Energy, Demand and Water Savings, ASHRAE Guideline, American Society of Heating, Refrigerating, and Air-Conditioning Engineers, 2014.
[60] J. Luo, T. Hong, S.-C. Fang, Benchmarking robustness of load forecasting models under data integrity attacks, Int. J. Forecast. 34 (1) (2018) 89–104, https://doi.org/10.1016/j.ijforecast.2017.08.004.
[61] M.R. Baker, K.H. Jihad, H. Al-Bayaty, A. Ghareeb, H. Ali, J.-K. Choi, Q. Sun, Uncertainty management in electricity demand forecasting with machine learning and ensemble learning: case studies of COVID-19 in the US metropolitans, Eng. Appl. Artif. Intell. 123 (2023) 106350, https://doi.org/10.1016/j.engappai.2023.106350.