Information Sciences 654 (2024) 119879

Contents lists available at ScienceDirect

Information Sciences

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

Baseline energy modeling for improved measurement and verification through the use of ensemble artificial intelligence models

Elissaios Sarmas *, Aikaterini Forouli, Vangelis Marinakis, Haris Doukas

Decision Support Systems Laboratory, School of Electrical & Computer Engineering, National Technical University of Athens, Greece

A R T I C L E I N F O

Keywords:
Measurement & verification
Baseline energy modeling
Energy efficiency
Explainable machine learning
Ensembling

A B S T R A C T

Accurate estimation of energy savings is crucial for the effective implementation of energy conservation measures (ECMs). Simultaneously, the integration of Artificial Intelligence (AI) has revolutionized software engineering by imbuing software with intelligent capabilities and autonomy. In this paper, we propose an ensemble model for precisely estimating baseline energy consumption within the realm of AI-empowered autonomous software. The ensemble model combines predictions from three tree-based Machine Learning models, namely Random Forest, XGBoost, and LightGBM. Notably, our model emphasizes the provision of explainability, granting transparency and insights into the key factors influencing baseline energy consumption. To validate its effectiveness, we conduct experimental evaluations on a diverse cluster of real-world buildings in Latvia. The results demonstrate the superiority of our proposed ensemble model over individual models and even a deep learning network tailored for energy consumption estimation. These findings underscore the efficacy of AI-empowered models in the energy sector, offering a robust and interpretable solution for estimating energy savings.

1. Introduction

Notwithstanding its significant share in global total final energy consumption (35%) and energy-related CO_2 emissions (38%) in 2019, the building sector has consistently maintained a stable level of global energy consumption and emissions since 2018 [19]. However, the operational energy consumption of buildings on a global scale still reached approximately 130 EJ, constituting roughly 30% of total final consumption, with electricity consumption in buildings now accounting for approximately 55% of global electricity consumption [18]. In light of these statistics, the reduction of energy consumption and emissions within the building sector emerges as a pivotal challenge in tackling climate change. Within this context, the implementation of building renovation measures emerges as a promising avenue for achieving these objectives [45]. Retrofitting existing buildings with energy-efficient systems, such as enhanced insulation and lighting, holds the potential for substantial energy savings and decreased greenhouse gas emissions. Nevertheless, despite the prospective benefits, the progress of building renovation has been sluggish, resulting in a substantial portion of the building stock remaining energy inefficient. Moreover, the COVID-19 pandemic has underscored the critical role of healthy and sustainable buildings, thereby amplifying the urgency for action in this domain. Governments and policymakers play a crucial role in incentivizing building renovation through the implementation of regulations, funding mechanisms, and other policy instruments [5].

* Corresponding author.
E-mail address: esarmas@epu.ntua.gr (E. Sarmas).

https://doi.org/10.1016/j.ins.2023.119879
Received 23 June 2023; Received in revised form 14 September 2023; Accepted 1 November 2023

Available online 8 November 2023
0020-0255/

Such endeavors can play an instrumental role in unleashing the potential of building renovation as a central strategy for diminishing energy consumption, mitigating climate change, and establishing a healthier and more sustainable built environment for the future [19].

ECMs constitute a fundamental instrument for diminishing energy consumption and greenhouse gas emissions across various sectors, including buildings, industry, and transportation [28]. These measures encompass a range of interventions, such as equipment upgrades, enhancements to insulation, lighting systems, and the implementation of energy-efficient practices. However, ensuring the efficacy of ECMs necessitates the precise measurement and verification (M&V) of the resulting energy savings. The significance of M&V has grown substantially in response to energy policies enacted worldwide. The European Union's Energy Efficiency Directive, for instance, mandates member states to achieve individual advancements in energy efficiency, thereby necessitating accurate and dependable estimation of energy savings derived from diverse ECMs to effectively execute the directive [4]. The extent to which energy savings can be reliably quantified relies upon the presence of uncertainty, rendering uncertainty analysis an indispensable step in the accurate estimation of energy savings.

In order to assess the energy savings resulting from an ECM, a comparative analysis must be conducted between the energy consumption during the reporting period (post-ECM) and the hypothetical consumption that would have taken place in the absence of the ECM. This comparison is commonly known as the adjusted baseline. Consequently, it becomes necessary to standardize the post-ECM consumption to align with the pre-ECM conditions [33]. The development of a baseline model utilizing engineering or statistical methodologies is typically employed to facilitate this normalization process. Moreover, to ensure the reliability and validity of the estimated energy savings, it is crucial to incorporate a quantifiable measure of uncertainty, which serves as an indicator of the accuracy associated with the savings estimation [14,39].

Precisely quantifying energy savings presents a significant challenge owing to the intricacies involved, necessitating the consideration of multiple variables and inherent uncertainties. As a result, the M&V of energy savings have garnered increasing attention in recent years, propelled by the aspiration to improve the dependability and accuracy of energy savings estimates. The crux of the challenge resides in the development of comprehensive models that encompass a broad spectrum of variables influencing energy consumption, spanning climatic conditions, occupancy patterns, and the performance of individual energy systems. The advent of digitization in the realm of buildings, accompanied by the proliferation of Internet of Things (IoT) devices that capture extensive datasets, holds great potential for advancing the development of sophisticated machine learning (ML) models for energy savings estimation [44,38,15]. These ML models possess the ability to integrate a wide spectrum of variables and data sources, thereby enabling more accurate and dependable energy savings estimates. The adoption of ML models for energy savings estimation presents a captivating opportunity to bolster the efficacy and outcomes of ECMs, thereby facilitating the achievement of ambitious energy policy objectives [2]. Nonetheless, the successful deployment of such models necessitates access to high-quality data and the expertise of proficient professionals well-versed in data analytics and modeling techniques.

This study aims to tackle the challenges pertaining to accurate energy savings estimation through a comparative analysis of diverse ML models employed in baseline energy modeling. The novelty of this study lies in the utilization of ML models, which aim to fully exploit the wealth of high-quality big data available for characterizing energy consumption in buildings. Specifically, the investigation aims to discern whether the implementation of ensembles comprising these ML models can further augment the accuracy of energy savings estimation. Furthermore, this study conducts extensive testing on a substantial cluster of real-world buildings located in Latvia, thereby offering valuable insights into the generalizability of the proposed models to practical applications. By delving into the potential of DL models and ensembles for energy savings estimation, this research endeavor contributes to the advancement of more efficacious ECMs, which are indispensably instrumental in attaining today's ambitious energy policy targets.

The contributions of our study are summarized as follows:

- The study proposes and compares different ML models for estimating baseline energy consumption in buildings, utilizing high-quality big data available from IoT devices.
- The study investigates the potential of ensembles of ML models for further increasing the accuracy of energy consumption estimation.
- The proposed models are tested on a large cluster of real-world buildings in Latvia, providing valuable insights into the effectiveness and generalizability of the models in practical applications.

The rest of the study is organized as follows. In section 2 the problem setting and literature review are presented. In section 3 the developed methodology is described. In section 4 an experimental application based on a real case study in Latvia is presented. Finally, concluding remarks are provided in section 5.

2. Problem setting and literature review

2.1. Baseline energy modeling for measurement & verification

Legislative efforts have predominantly focused on the implementation of ECMs, with the aim of reducing energy consumption. The effectiveness of this endeavor hinges heavily on the accuracy of M&V conducted for each individual ECM, ensuring a reliable estimation of energy savings. However, a significant concern arises regarding the potential overestimation of savings at the project level, which could hinder climate change mitigation efforts. Consequently, there exists an urgent need for a methodology capable of surmounting the barriers that impede accurate M&V in building facilities, encompassing challenges related to cost, resource

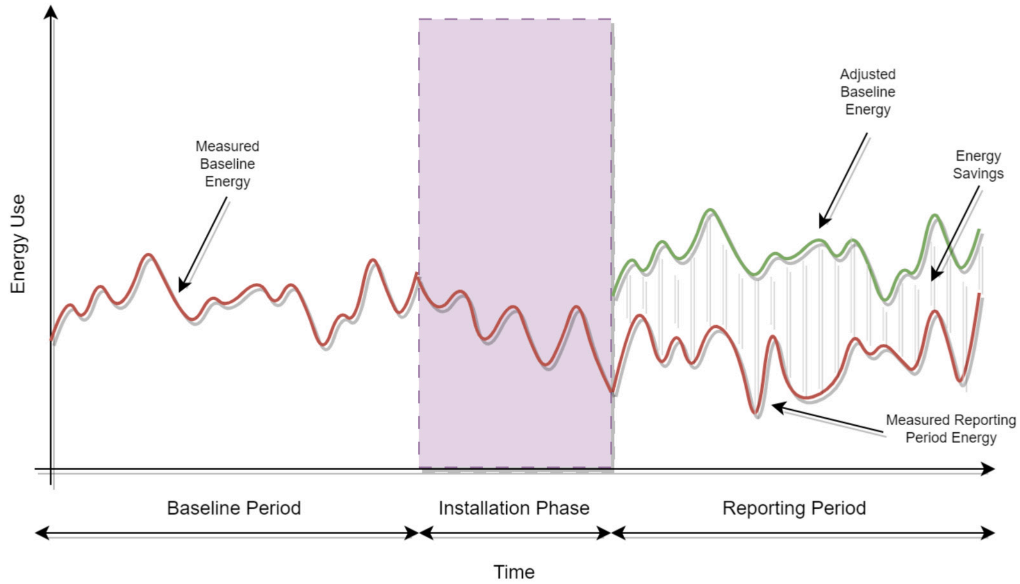

Fig. 1. Energy savings calculation according to the IPMVP framework. Savings can be estimated with measurements before and after the ECM installation, but a normalisation of the baseline energy is also required to incorporate weather and other effects.

limitations, and time constraints. Previous research has recognized the potential of ML as a viable tool to address these challenges, highlighting its suitability in this context [13].

In the context of M&V projects, three distinct periods of significance are evident, namely the baseline, implementation, and reporting periods. These periods are sequentially organized, yet their duration varies depending on specific project parameters. Preceding the implementation of ECMs, the baseline period establishes a reference point, while the reporting period ensues after ECM implementation. A critical task within the M&V process involves the estimation of the adjusted baseline during the reporting period by normalizing energy consumption to baseline conditions. Typically, engineering or statistical methodologies are employed to construct a baseline model capable of achieving this normalization. Consequently, it should be recognized that M&V is an inherently imprecise discipline, and the maintenance of accuracy throughout the process is paramount to its success. The primary sources of uncertainty in M&V encompass sampling, measurement, and modeling. To address this, the International Performance Measurement and Verification Protocol (IPMVP) delineates a methodological framework for quantifying uncertainty in a project, with the minimum acceptable level of uncertainty defined as the threshold where savings surpass twice the standard error of the baseline value.

The IPMVP serves as the predominant and globally adopted methodology for M&V, comprising four distinct approaches that cater to a wide array of retrofitting actions (Fig. 1). Options A and B establish a project boundary that isolates the retrofit and encompasses the affected equipment. Option C adopts a comprehensive whole-building approach, suitable for scenarios where savings surpass 10% of the total site energy consumption. Option D involves a calibrated simulation of energy systems, proving advantageous in situations where baseline data is unavailable. Complementary to the IPMVP, ASHRAE Guideline 14 and ISO50015 offer methodologies grounded in its core principles, providing robust frameworks. Nonetheless, a notable drawback of these generalized methodologies is the acknowledged lack of guidance concerning the calculation process.

The determination of savings can encompass either the entire facility or a specific portion thereof, contingent upon the characteristics of the Energy Efficiency Measure(s) (EEMs) and the intended purpose of reporting. To isolate the equipment and its associated energy usage affected by the EEMs from those unaffected, a measurement boundary is established. Within this boundary, all energy consumed or generated must be measured or estimated using meters. It is crucial to consider all energy flows across the measurement boundary, including instances where energy may flow in reverse, such as with on-site solar generation. The selection of a measurement boundary aligns with one or more of the four IPMVP Options, which impact the level of detail in reported savings and the required measurements. The choice of Option(s) should consider the purpose(s) of the M&V reporting.

If the reporting objective is to verify the savings resulting from the EEMs' impact on equipment, a measurement boundary should encompass that specific equipment, allowing for the determination of measurement requirements within the boundary. This approach corresponds to the Retrofit Isolation Option (Option A or B). If the reporting goal involves verifying and/or managing the overall energy performance of the entire facility or validating savings resulting from multiple EEMs with interactive effects, the meters measuring the energy supply to the entire facility can be utilized to assess performance and savings. In this case, the measurement boundary encompasses the entire facility, and the approach employed is Option C: Whole Facility. In situations where unreliable or unavailable data exist for the Baseline Period or Reporting Period (e.g., new construction projects), calibrated simulation models can be employed to provide energy data for either a portion or the entirety of the facility. The measurement boundary can be delineated accordingly, and the approach utilized is Option D: Calibrated Simulation.

Table 1
Literature review on baseline energy modeling studies.

#	Author (Year)	Models Used	Metrics Used	Building Type
1	[43]	GBM	R2, CV(RMSE)	Commercial
2	[16]	GBM	CV(RMSE)	Commercial
3	[26]	iTrAdaBoost	MAPE, RMSE, MAE	Educational
4	[22]	SVM	RMSE, MAE	Industrial
5	[1]	LR, OLS, KNN, SVM	MASE, MAPB	Commercial
6	[13]	OLS, DT, KNN, SVM, ANN	med(absRTE), CV(RMSE), NMBE	Industrial
7	[30]	NARX-ANN	MSE, RMSE	Educational
8	[49]	LR, ANN	MSE	Residential
9	[14]	KNN, DT, RF	NMBE, CV(RMSE)	Multiple
10	[12]	ANN	SSE, MSE	Multiple
11	[48]	ANN	CV(RMSE)	Commercial
12	[27]	BPNN, RBFNN, GRNN, SVM	RMSE, MRE	Commercial
13	[32]	ANN, EnergyPlus	SE	Commercial
14	[23]	ANN	CV(RMSE), MBE	Commercial
15	[37]	ANN	R2, RMSE, MAE	CHP Plant
16	[10]	Gaussian Mixture	R2	Commercial

2.2. Related work

Several approaches have been tested in the literature to model buildings' consumption in the baseline period. It is evident that ML-based energy modeling approaches outperform traditional linear regression approaches [49,30] and their suitability is proven in a wide range of studies in commercial [14,12,48,27,32,23,17], residential [11,20,1] as well as industrial [13,25,37] applications. On finding the optimal ML approach, and taking into consideration the diverse impact of characteristics (like selected input variables, amount and quality of training available data, granularity of data and forecasts etc.) on the model performance, it is important to compare the different modeling techniques within the context of the unique project conditions and under same circumstances.

Under this perspective, there are a lot of studies that compare different ML techniques in the context of the same application. [13] applied six ML techniques, namely bi-variable and multi-variable ordinary least squares regression, decision trees, k-nearest neighbors, artificial neural networks and support vector machines, for M&V of energy savings in the context of industrial buildings (biomedical manufacturing facility). The optimal model was a single layer feed-forward neural network trained using monthly data, showing a 51% reduction in error compared to conventional approaches. [43] proposed a gradient boosting machine and assessed its performance in a large set of commercial buildings against a piece-wise linear regression and a random forest (RF) algorithm. The new algorithm showed an improvement of the prediction accuracy in more than 80% of the test cases. The study of [27] compared the traditional back propagation neural network (BPNN), the radial basis function neural network (RBFNN), general regression neural network (GRNN) and support vector machine (SVM) methods to predict the cooling load of an office building in China. The results showed that the accuracy of the SVM and GRNN methods were better than the BPNN and RBFNN methods. Finally, in the paper of [1], eleven linear and non linear algorithms are compared, four of them being ML algorithms, namely k-nearest neighbors, support vector machines regression, locally-weighted smoothing splines, tree-based algorithm cubist and RFs. In this application, the performance of each method varied between the different types of buildings and there is no algorithm that clearly outperformed the others. An exhaustive list of relevant literature is presented in Table 1.

The accuracy of savings estimations is also dependent on the selection of the model input variables. Factors that most influence the buildings' energy consumption should be selected and incorporated with the most appropriate metrics in the baseline models. These factors may relate to climate and weather conditions, such as humidity, cooling or heat degree days, wind speed, solar radiation [10,22], the type and intensity of usage including occupancy levels, type of building, number of working days per month, building schedule - production level, operating schedule [22], the time period of use depending on the model time granularity (season, month of year, weekday/ weekend, hour of day), the building characteristics, including building global heat loss coefficient, orientation, insulation thickness [12] or any other interconnected set of the aforementioned variables. Each of the above factors may be translated into model variables in different ways.

We refer indicatively to outdoor air humidity, dew point or wet bulb temperature, and enthalpy as alternative ways to model humidity [36]. Each of the above variables could also interact with each other and lead to the introduction of new climatic variables, like "moist air enthalpy", when mixing temperature and humidity, and "potential evapo-transpiration" proposed by [1] as a function of air pressure, humidity, temperature, wind speed, solar radiation, and others. [1] also proposed an automated selection method to identify the best variable sets for the analyzed buildings. Novel model features are also introduced in the work of [8], who used the building global heat loss coefficient, the south equivalent surface, and the difference in the indoor set point temperature and the solar temperature to predict heating energy demand.

Other parameters that affect prediction performance are the model training period and the measurements and forecasts frequency. [13] investigated how the training period characteristics could impact the model's performance. It was found that the use of a higher measurement frequency reduced the spread of error across the examined models. However, further analysis proved the use of more granular data did not always benefit model performance. In the analysis of [20] for multi-family residential buildings, in a frequency scale from 10 min to daily measurements, an hourly granularity provided the best results in terms of prediction performance. The paper of [1] also examined models with different frequency of measurements (monthly and daily consumption values are used), but these were applied in different buildings and cannot be compared. At any case, in selecting the model parameters and configuration, it is important to consider that in real practical scenarios, the quality of available datasets, their sampling frequency and resolution may differ.

3. Methodology for modeling baseline energy

In this section we introduce the proposed methodology for modeling baseline energy. Firstly, the overall methodological framework is presented by highlighting the need for ensembling different ML models which can lead to better forecasts. Secondly, the selected ML models (RF, XGBoost and LightGBM) for predicting baseline energy are presented.

3.1. Methodology overview

Within the domain of forecasting, it is widely acknowledged that no single model can consistently surpass all available models in terms of predictive performance [34]. This notion holds true for the domain of M&V and baseline energy modeling, as well. Consequently, forecasters have resorted to combining the predictions generated by multiple models, each employing distinct assumptions and possessing a different structural configuration, in an endeavor to enhance overall forecast accuracy. The underlying principle behind forecast combinations resides in the expectation that the errors inherent in diverse models will offset one another, resulting in more accurate forecasts than those produced by individual forecasting approaches. Numerous schemes for combining forecasts have been proposed, encompassing both simplistic methods such as averaging (e.g., mean, median, or mode) and more sophisticated techniques (e.g., leveraging in-sample and out-of-sample forecasting errors, information criteria, Bayesian probability theory, and linear or nonlinear regression methodologies).

In the context of our research, investigating baseline energy prediction subsequent to the implementation of ECMs, we have devised a stacking ensemble model. The proposed ensemble model consists of three ML-based models, namely RF, XGBoost, and LightGBM. Each of these models possesses distinct strengths and characteristics, capturing diverse aspects of the underlying energy consumption patterns. By harnessing the collective predictive power of these models, our objective is to optimize the accuracy of baseline energy predictions. The stacking ensemble model operates by aggregating the predictions generated by the individual ML models. However, to further enhance the fusion process, we introduce a filtering step employing an average-based ensembling model, as shown in Fig. 2. This supplementary layer facilitates an optimal amalgamation of the predictions derived from the base models, capitalizing on the unique strengths exhibited by each model. Through this iterative procedure, the stacking ensemble model harnesses the complementary capabilities of the individual ML models, resulting in an ensemble prediction characterized by improved accuracy and robustness. The adoption of this stacking ensemble approach in our study constitutes a noteworthy contribution to the domain of M&V and baseline energy modeling. By effectively integrating the expertise and insights derived from different ML models, the proposed model aims to address the limitations inherent in individual models, thereby enhancing the precision of energy savings estimation.

3.2. Baseline models

3.2.1. Random forest

The Random Forest (RF) algorithm stands as an immensely popular and extensively employed technique within the realm of data science [7]. It represents a supervised ML algorithm that finds its applications in a multitude of classification and regression problems. Through the construction of decision trees on distinct samples and the subsequent aggregation of their predictions via majority voting or averaging, the RF algorithm attains exceptional levels of accuracy.

A pivotal strength ingrained within the RF lies in its capacity to tackle datasets comprising both continuous and categorical variables. This innate versatility renders it highly suitable for a wide array of classification and regression tasks. In the forthcoming tutorial, we shall embark on an exploration of the intricate inner workings of the RF algorithm, simultaneously demonstrating its implementation on a compelling classification problem.

At the heart of the RF algorithm resides ensemble learning, a concept instrumental to its functioning. Ensemble learning amalgamates multiple models, eschewing reliance on a solitary model [40]. Within the confines of the RF algorithm, this ensemble technique adopts the principles of bagging [6]. Accordingly, it fashions various training subsets from the original dataset, employing sampling with replacement. Each of these subsets engenders the training of an independent decision tree model, culminating in a final prediction realized through the aggregation of all the decision tree predictions [35].

RF bestows numerous advantages upon its practitioners. It exhibits prowess in effectively handling non-linear relationships and noisy datasets [42]. Furthermore, it boasts a diminutive requirement for data pre-processing, distinguishing itself from certain alternative algorithms. It is important, however, to acknowledge that RF sacrifices the interpretability of individual trees in favor of bolstered predictive accuracy, while due to the algorithm's intricacies, very extensive parameter tuning is imperative.

3.2.2. XGBoost

Gradient boosting (GB) emerges as a versatile ensemble learning algorithm, catering to both regression and classification tasks [31]. Its fundamental objective revolves around the construction of a “strong learner” model through the iterative amalgamation of multiple “weak learner” models. Each subsequent model aims to rectify the prediction errors of its predecessor [41]. Following a methodology akin to other boosting techniques, GB employs decision trees (regressors or classifiers) trained on the negative gradient

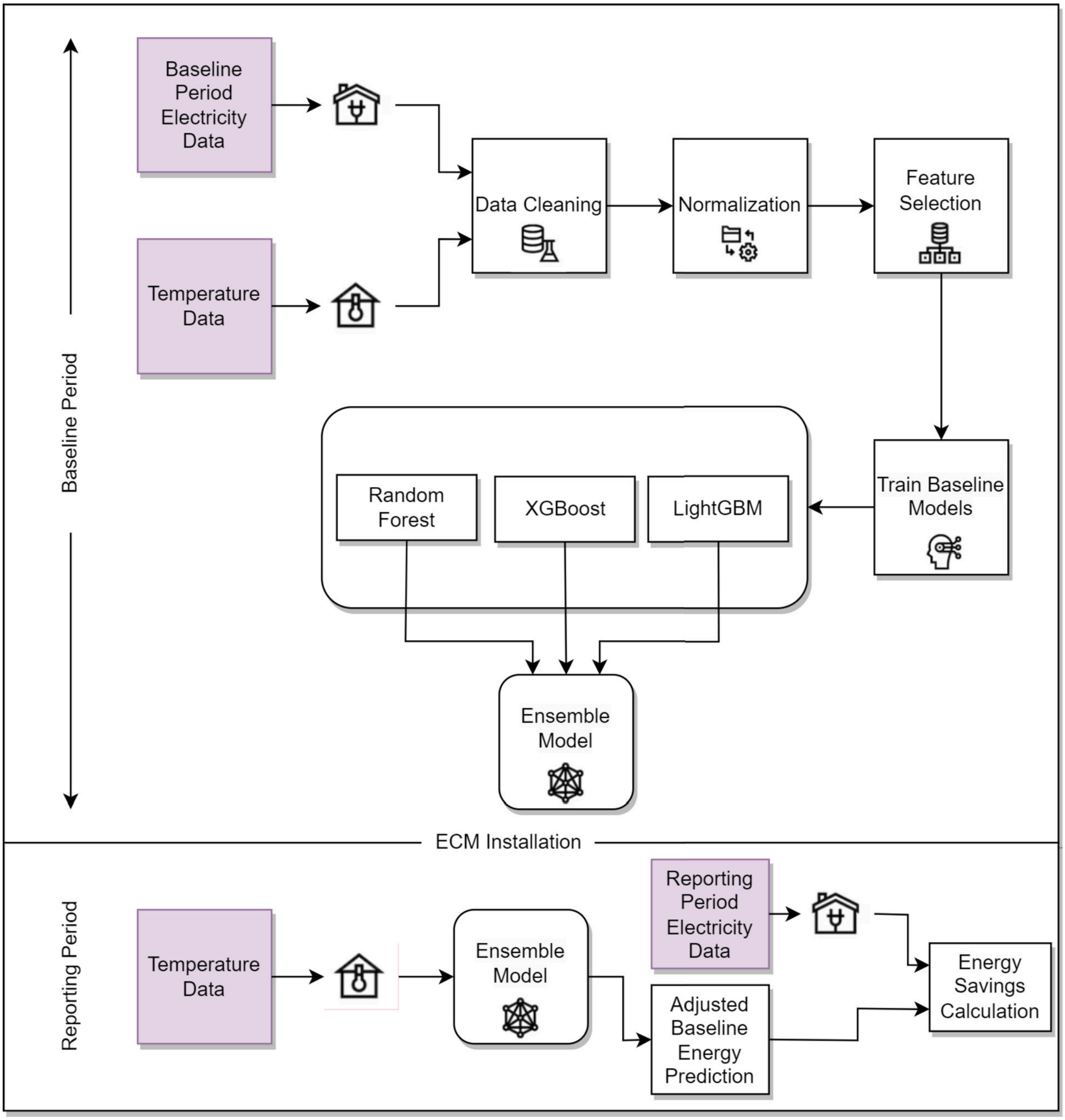

Fig. 2. The proposed methodological framework for estimating energy savings based on ML models and ensembling.

of the loss function at each stage. This approach enables the newly added models to address the specific deficiencies of prior models, resulting in an overall performance improvement. Once all individual trees are integrated into the ensemble, the final model takes charge of predicting new samples.

XGBoost, an optimized variant of GB, belongs to the GB family but distinguishes itself through the incorporation of advanced regularization techniques, namely $L1$ and $L2$ regularization, augmenting the generalization capabilities of the resulting models. Functioning as an efficient implementation of the GB algorithm, XGBoost offers superior computational performance compared to traditional GB models. Additionally, it reduces training times through parallelization across clusters [9]. Similar to RF, XGBoost also assembles ensembles from decision tree models. However, the primary divergence lies in the sequential addition of trees to the ensemble, with the aim of minimizing prediction errors from prior models (this distinction also represents the fundamental disparity between bagging and boosting). The term “gradient boosting” stems from the adoption of the gradient descent optimization algorithm to minimize errors.

XGBoost boasts several advantages that have solidified its status as one of the foremost models for GB. As mentioned earlier, it employs a tailored pruning technique for decision trees, resulting in expedited training and enhanced handling of voluminous datasets [46]. Among its optimizations, XGBoost leverages an approximate greedy algorithm of weighted quantiles during node splitting, bypassing the evaluation of all possible splits. Moreover, it partitions data into smaller segments, facilitating parallel processing of samples, while harnessing cache memory to accelerate calculations [47]. These features have contributed to its widespread adoption across diverse classification and regression applications over the past decade.

Analogous to RF, XGBoost relies on a crucial set of hyperparameters to achieve high-accuracy predictions. These hyperparameters encompass aspects not only related to the GB process, such as the number of trees utilized during training but also pertaining to the structure of each individual tree. Notable parameters include tree depth and the minimum loss reduction necessary for further node partitioning. For this study, the original XGBoost implementation library was employed.

3.2.3. LightGBM

LightGBM, introduced by [24], is a notable GB framework similar to XGBoost. While both are based on the GB algorithm and use decision trees, LightGBM prioritizes performance, scalability, and memory efficiency. It diverges from XGBoost in decision tree construction, using a leaf-wise strategy that selects the leaf with the highest loss decrease. LightGBM also employs a histogram-based learning algorithm to determine optimal split points, improving efficiency and reducing memory requirements. This framework has been developed by Microsoft researchers, is a framework and it combines decision trees sequentially, where each new tree improves the model by fitting the residuals from the previous tree. LightGBM has achieved remarkable accuracy in ML competitions [29].

The motivation behind LightGBM stems from the process of building decision trees, which involves splitting observations based on feature values to reduce randomness and increase information gain. Finding the best split, which results in the highest information gain, is the most time-consuming aspect of tree learning [21]. Other gradient boosting implementations utilize two algorithms to find the best splits: pre-sorted and histogram-based. The pre-sorted algorithm involves sorting feature values and evaluating all possible split points, while the histogram-based algorithm discretizes continuous features into bins to create feature histograms. LightGBM aims to address the efficiency issue, especially with large datasets.

The motivation behind LightGBM is to improve the efficiency of building decision trees. The process involves splitting observations based on feature values to reduce randomness and increase information gain. Finding the best split, which maximizes information gain, is time-consuming. Other gradient boosting decision tree implementations use pre-sorted or histogram-based algorithms to find splits. LightGBM addresses efficiency issues, especially with large datasets [3]. To enhance efficiency, LightGBM incorporates two techniques: Gradient One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS prioritizes instances with larger gradients to accelerate information gain, while EFB reduces the number of features by grouping mutually exclusive features. These techniques make LightGBM efficient for ranking, classification, and regression tasks.

4. Case study

This section provides an extensive experimental application of the proposed methodology on data stemming from real renovation projects. The training process of the baseline classification methods is described in detail, including the fine-tuning of their hyperparameters, followed by the evaluation of the meta-learning model. Finally, the results of the application are presented and discussed.

4.1. Data set

In this subsection, we will introduce the dataset that was utilized to apply the proposed methodology in an experimental setting. The evaluation encompassed a total of 8 actual buildings, and we will delve into the accuracy metric of the models in the subsequent subsection. For the purposes of this section, our focus will be on analyzing the dataset of a specific building: a public library located in the town of Gulbene, situated in north-eastern Latvia. It is important to note that similar renovations have been carried out in the remaining 7 buildings. These renovations comprised exterior wall insulation, cap insulation, modernization of the drainage system, replacement of doors and windows, and restoration of heating system pipe insulation.

The dataset consists of 1.5 years' worth of electricity consumption data, collected with an hourly frequency using smart meters installed in the building. Additionally, hourly weather data were obtained from a meteorological station in Aluksne, which is a nearby town to Gulbene. To ensure the quality and reliability of the dataset, a preprocessing step was undertaken. This involved removing outliers and addressing missing data. In the case of missing data, we applied linear interpolation, replacing the missing values with interpolated values based on data from the same hour of the previous and next days.

The dataset provides a solid foundation for analyzing energy consumption patterns and examining their correlation with temperature fluctuations. Fig. 3 provides a visual representation of the temporal variation in temperature and electricity consumption. The upper part of the figure showcases the temperature values, while the lower part displays the electricity consumption. This plot allows us to discern any noticeable patterns, trends, or correlations that may exist between these two variables.

Fig. 4 illustrates the combined distribution of temperature and electricity consumption in a single plot. The frequency of temperature values is represented on the left y-axis, while the frequency of electricity consumption values is shown on the right y-axis. The x-axis displays the corresponding ranges of temperature and electricity. This histogram visualization enables us to examine the distribution of the data and detect any prominent peaks or patterns.

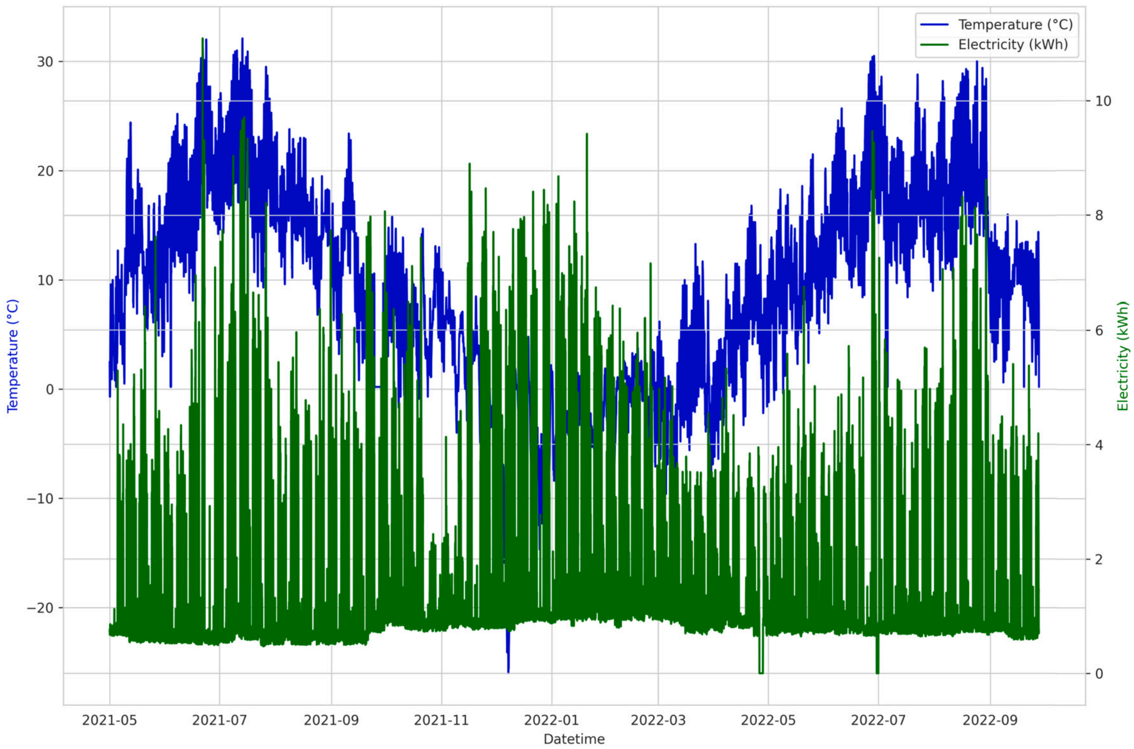

Fig. 3. Time-series of temperature (blue color) and electricity consumption (green color) over time.

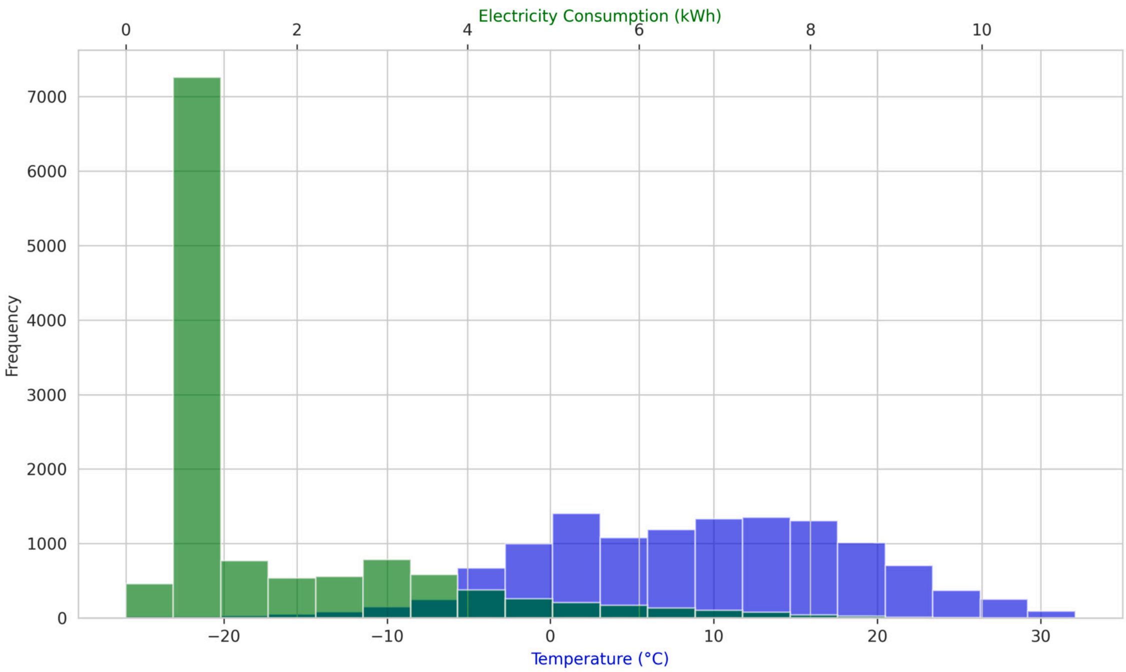

Fig. 4. Histogram of Temperature and Electricity showcasing the distribution of temperature and electricity time-series for a given building.

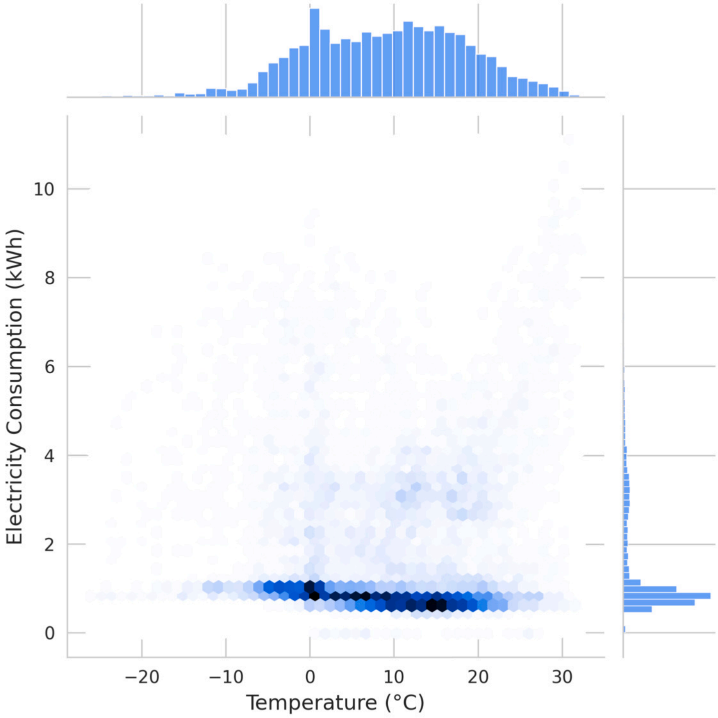

Fig. 5. Jointplot of Temperature and Electricity highlighting the correlation between the two time-series.

Fig. 5 depicts the correlation between temperature and electricity consumption through a jointplot. This plot combines a scatter plot and a histogram to offer a comprehensive view of the joint distribution of these variables. The hexagonal bins portray the density of data points, with the color indicating the level of density. By examining this plot, we can gain insights into the potential correlations or relationships between temperature and electricity consumption.

Lastly, it is evident that the electricity consumption time-series exhibits significant autocorrelation. This observation is supported by Fig. 6, which presents the seasonal decomposition of the electricity time-series data. The decomposition reveals distinct components of the electricity consumption: trend, seasonality, and residuals. The upper sub-plot represents the observed electricity values, while the second sub-plot illustrates the trend component. The lower sub-plots display the seasonal and residual components. This decomposition enhances our comprehension of the inherent patterns and fluctuations in the electricity consumption data, confirming the presence of pronounced daily and weekly autocorrelation.

4.2. Experimental design

In this subsection, we describe the training process of the electricity consumption forecasting model and provide an overview of its features. The dataset used for training and testing purposes consists of 1.5 years of electricity consumption data, with one year allocated for model training and the remaining data reserved for testing.

The electricity consumption forecasting model is designed to provide hourly predictions for the energy consumption of the selected target building. The choice for hourly output data was made to ensure aligned point forecasts for the adjusted baseline energy and the reporting period measured energy, leveraging the reliable hourly measurements provided by smart meters. The model incorporates weather and seasonal features, specifically including outdoor air temperature, hour of the day, day of the week, weekday or weekend (binary variable), month of the year, and vacation period indicator (binary variable). The consideration of vacation periods is particularly important for office buildings, as they tend to have lower energy consumption during these periods. For this study, August and Christmas were considered as the only vacation periods.

The developed forecasting models are evaluated using three error metrics: the Mean Absolute Error (MAE), the Coefficient of Variation of the Root Mean Squared Error (CVRMSE), and the Normalized Mean Bias Error (NMBE). MAE measures the individual differences between the real and predicted values, and is calculated as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

CVRMSE, recommended by ASHRAE 14 (Ashrae, 2020), is a normalized version of the Root Mean Squared Error (RMSE) index. It is calculated as the RMSE divided by the average consumption value, and is given by the following equation:

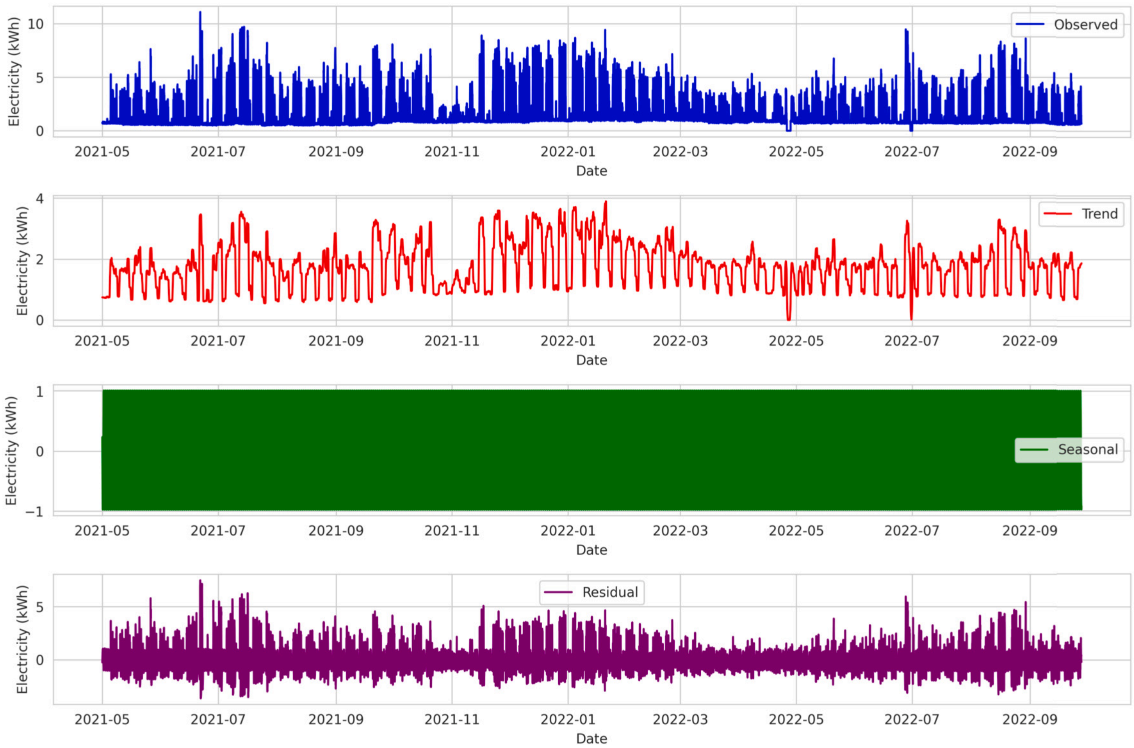

Fig. 6. Seasonal Decomposition of the Electricity consumption time-series.

$$\mathrm{CVRMSE} = \frac{1}{\bar{y}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

The Normalized Mean Bias Error (NMBE) measures the mean bias between the real and predicted values, normalized by the average consumption value. It is calculated as follows:

$$\mathrm{NMBE} = \frac{1}{\bar{y}}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)$$

These error metrics provide valuable insights into the accuracy and performance of the forecasting models, enabling us to assess the quality of the predictions and compare different approaches.

Finally, fine-tuning ML models is an indispensable step towards achieving optimal performance. In this section, we delve into the fine-tuning process and discuss the specific hyperparameters chosen for RF, XGBoost, and LightGBM.

For RF, the number of estimators was set to 200 to strike a balance between model complexity and computational resources. A maximum depth of 10 was meticulously chosen to capture intricate functions without succumbing to overfitting. To promote enhanced generalization, the maximum features hyperparameter was configured as “sqrt”, considering the square root of the total number of features. This approach allows the model to avoid undue reliance on any single feature. Bootstrapping was enabled by setting the bootstrap hyperparameter to True, introducing diversity in the training data and improving model robustness.

In the case of XGBoost, 500 estimators were employed to provide the model with greater opportunities to learn complex patterns. A maximum depth of 6 was chosen to strike a balance between capturing complex relationships and preventing overfitting. The learning rate was set to 0.1 to facilitate meticulous weight updates, leading to better convergence and improved generalization.

For LightGBM, 300 estimators were used to enable the model to learn intricate patterns, albeit with longer training times. A maximum depth of 8 was chosen to prevent overfitting, aligning with the best practices for RF and XGBoost. The number of leaves was set to 64, corresponding to $2^{\text{Maximum depth}}$, striking a balance between model complexity and overfitting. A learning rate of 0.05 was selected to facilitate cautious weight updates, promoting better convergence and reducing the risk of divergent behavior.

4.3. Results

The performance assessment involved evaluating five different models for electricity consumption forecasting in eight buildings located in Latvia. These models included an Artificial Neural Network (ANN) based on deep learning, as well as RF, XGBoost, LightGBM, and an ensemble model combining all tree-based individual models.

The experimental results are presented in Table 2. Notably, the ensemble model consistently outperformed RF, XGBoost, and LightGBM across all buildings. This comprehensive model exhibited superior predictive performance, as indicated by its consistently

Table 2
Building performance comparison using different forecasting models.

Building	Error Metric	ML Models				
		ANN	RF	XGBoost	LightGBM	Ensemble
Building 1	CV(RMSE)	0.10	0.12	0.13	0.11	0.10
	NMBE	-0.02	-0.01	-0.03	-0.02	-0.01
	MAE	5.25	5.28	5.23	5.27	5.23
Building 2	CV(RMSE)	0.11	0.12	0.11	0.10	0.10
	NMBE	-0.01	-0.02	-0.01	0.00	0.00
	MAE	4.92	4.95	4.88	4.91	4.91
Building 3	CV(RMSE)	**0.09**	0.11	0.12	0.10	**0.09**
	NMBE	-0.01	-0.03	-0.02	-0.01	-0.01
	MAE	6.03	6.08	**5.97**	6.02	6.02
Building 4	CV(RMSE)	0.14	0.15	0.16	0.14	**0.13**
	NMBE	0.01	0.02	0.03	0.01	**0.00**
	MAE	5.70	5.73	**5.65**	5.69	5.70
Building 5	CV(RMSE)	**0.12**	0.13	0.14	0.12	**0.12**
	NMBE	-0.02	**-0.01**	-0.03	-0.02	-0.02
	MAE	6.18	6.22	**6.15**	6.19	6.19
Building 6	CV(RMSE)	0.13	0.15	0.15	0.14	**0.12**
	NMBE	-0.02	-0.01	-0.03	-0.02	**-0.01**
	MAE	5.42	5.45	**5.38**	5.41	5.39
Building 7	CV(RMSE)	0.10	0.12	0.11	**0.09**	**0.09**
	NMBE	0.00	0.01	-0.01	0.00	**0.00**
	MAE	4.97	5.00	4.96	**4.95**	4.95
Building 8	CV(RMSE)	**0.08**	0.10	0.11	0.09	**0.08**
	NMBE	-0.01	-0.02	-0.03	-0.02	**-0.01**
	MAE	5.88	5.92	**5.84**	5.89	5.85

lower CV(RMSE) values. These findings underscore the efficacy of combining multiple models to enhance the accuracy of electricity consumption forecasting.

It is worth noting that the performance of the deep learning model exhibited variability depending on the individual buildings examined. While ANN demonstrated superior performance compared to RF, XGBoost, and LightGBM in certain cases, its results were similar or slightly inferior in others. This observation indicates that the efficacy of deep learning architectures, such as ANN, for electricity consumption forecasting might be influenced by the distinct characteristics of the building and dataset being analyzed.

These findings highlight the limitations of relying solely on deep learning for electricity consumption forecasting. It is evident that a more effective approach involves utilizing ensemble models that incorporate non-deep learning methods. By combining different models, ensembles offer a robust and dependable solution for accurate forecasting. This emphasizes the importance of considering a diverse range of models and tailoring the ensemble configuration to suit specific building scenarios.

Although deep learning models such as ANN have shown promise in certain instances, the ensemble model consistently outperformed both deep learning and other individual models. This underscores the significant value of ensembles in achieving precise and efficient energy consumption predictions. The results of this study provide valuable insights into model selection and configuration for electricity consumption forecasting, thereby contributing to improved energy management strategies and enhanced building energy efficiency.

5. Conclusions and future prospects

The estimation of energy savings from retrofitting actions in the building sector is very important for various reasons. The infusion of AI into software engineering has led to a paradigm shift, equipping software with intelligent capabilities and autonomy. Within this context, we have introduced an ensemble model designed to accurately estimate baseline energy consumption in AI-empowered autonomous software systems. The proposed ensemble approach combines multiple models to achieve enhanced accuracy and robustness in estimating energy savings. This underscores the importance of leveraging ensemble techniques to harness the predictive power of diverse algorithms and improve overall performance. The success of the ensemble model can be attributed to the complementary nature of its constituent models, which mitigates individual weaknesses and enhances overall predictive capability. By integrating the diverse perspectives and learning strategies of XGBoost, RF, and LightGBM, the ensemble model effectively captures complex patterns and relationships within energy data, resulting in more accurate estimations.

The proposed ensemble model not only excels in accuracy but also prioritizes the provision of explainability. This emphasis on transparency grants valuable insights into the key determinants of baseline energy consumption, setting our approach apart from prior works. When compared to earlier methodologies, our model offers a higher degree of trustworthiness, enabling stakeholders to comprehend and trust the energy savings estimates it generates. This can be attributed to the use of tree-based models that allow for a greater extent of explainability compared to deep neural networks which are very often employed in the last years. Thus future work can focus on examining which features finally drive the forecasts produced by the models.

Another interesting point would be to shed light on the untapped potential of personalized forecasts. These forecasts can be tailored to diverse building types, climate conditions, and other contextual variables, thus empowering stakeholders to make informed decisions that cater to specific needs and constraints. Future research could further investigate model development towards more personalized forecasts in order to optimize energy consumption and enhance overall energy efficiency, both in AI-empowered software and the broader energy sector. Regarding future research, it is essential to evaluate the robustness of our model under extreme conditions, such as extremely high energy demand or exceptionally low energy production. This resilience underscores its practical applicability and reinforces its status as a reliable and generalizable model for estimating energy savings, regardless of the prevailing conditions (i.e. in buildings located in countries with different climate conditions).

In addition, the standardization of data collection protocols is imperative to ensure consistency and comparability across studies. This involves harmonizing data formats, defining common variables, and establishing standardized M&V procedures. Standardization facilitates data sharing, collaborative methodologies, and result replication and validation among researchers and practitioners. Finally, achieving consensus on error metrics for evaluating energy savings estimation models is essential. While this study considered various error metrics such as CV(RMSE), NMBE, and MAE, it is crucial to establish a standardized set of metrics specifically tailored to energy savings estimation. This enables meaningful comparisons, benchmarking, and provides a common framework for assessing model accuracy, fostering the development of best practices in the field.

CRediT authorship contribution statement

Elissaios Sarmas: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft. **Aikaterini Forouli:** Formal analysis, Investigation, Methodology, Validation, Writing – original draft. **Vangelis Marinakis:** Conceptualization, Formal analysis, Project administration, Writing – original draft. **Haris Doukas:** Funding acquisition, Investigation, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

Data availability

Data will be made available on request.

Acknowledgements

The work presented is based on research conducted within the framework of the H2020 European Commission project BD4NRG (Grant Agreement No. 872613). The authors wish to thank Aija Zučika from the Latvian Environmental Investment Fund (LEIF), whose contribution, helpful remarks and fruitful observations were invaluable for the development of this work. The content of the paper is the sole responsibility of its authors and does not necessary reflect the views of the EC.

References

[1] M. Agenis-Nevers, Y. Wang, M. Dugachard, R. Salvazet, G. Becker, D. Chenu, Measurement and verification for multiple buildings: an innovative baseline model selection framework applied to real energy performance contracts, Energy Build. 249 (2021) 111183.
[2] R. Agrawal, L. De Tommasi, P. Lyons, S. Zanoni, G. Papagiannis, C. Karakosta, A. Papapostolou, A. Durand, L. Martinez, G. Fragidis, et al., Challenges and opportunities for improving energy efficiency in smes: learnings from seven European projects, Energy Effic. 16 (2023) 17.
[3] E. Al Daoud, Comparison between xgboost, lightgbm and catboost using a home credit dataset, Int. J. Comput. Inf. Eng. 13 (2019) 6–10.
[4] C.A. Balaras, E.G. Dascalaki, K.G. Droutsa, S. Kontoyiannidis, Empirical assessment of calculated and actual heating energy use in hellenic residential buildings, Appl. Energy 164 (2016) 115–132.
[5] P. Bertoldi, M. Economidou, V. Palermo, B. Boza-Kiss, V. Todeschi, How to finance energy renovation of residential buildings: review of current and emerging financing instruments in the eu, Wiley Interdiscip. Rev., Energy Environ. 10 (2021) e384.
[6] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.
[7] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[8] T. Catalina, V. Iordache, B. Caracaleanu, Multiple regression model for fast prediction of the heating energy demand, Energy Build. 57 (2013) 302–312.
[9] 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 KDD '16, ACM, New York, NY, USA, 2016, pp. 785–794.
[10] J.A. Díaz, J.S. Ramos, M.C.G. Delgado, D.H. García, F.G. Montoya, S.Á. Domínguez, A daily baseline model based on transfer functions for the verification of energy saving. A case study of the administration room at the palacio de la madraza, granada, Appl. Energy 224 (2018) 538–549.
[11] R.E. Edwards, J. New, L.E. Parker, Predicting future hourly residential electrical consumption: a machine learning case study, Energy Build. 49 (2012) 591–603.
[12] B.B. Ekici, U.T. Aksoy, Prediction of building energy consumption by using artificial neural networks, Adv. Eng. Softw. 40 (2009) 356–362.

[13] C.V. Gallagher, K. Bruton, K. Leahy, D.T. O'Sullivan, The suitability of machine learning to minimise uncertainty in the measurement and verification of energy savings, Energy Build. 158 (2018) 647–655.
[14] J. Granderson, S. Touzani, C. Custodio, M.D. Sohn, D. Jump, S. Fernandes, Accuracy of automated measurement and verification (m&v) techniques for energy savings in commercial buildings, Appl. Energy 173 (2016) 296–308.
[15] C. Granell, A. Kamilaris, A. Kotsev, F.O. Ostermann, S. Trilles, Internet of things, in: Manual of Digital Earth, 2020, pp. 387–423.
[16] B. Grillone, G. Mor, S. Danov, J. Cipriano, A. Sumper, A data-driven methodology for enhanced measurement and verification of energy efficiency savings in commercial buildings, Appl. Energy 301 (2021) 117502.
[17] J.S. Haberl, S. Thamilseran, Great Energy Predictor Shootout ii: Measuring Retrofit Savings–Overview and Discussion of Results, 1996.
[18] I. IEA, Energy Technology Perspectives 2020, International Energy Agency, 2020.
[19] U. IEA, 2020 global status report for buildings and construction, Tech. Rep., 2020.
[20] R.K. Jain, K.M. Smith, P.J. Culligan, J.E. Taylor, Forecasting energy consumption of multi-family residential buildings using support vector regression: investigating the impact of temporal and spatial monitoring granularity on performance accuracy, Appl. Energy 123 (2014) 168–178.
[21] D. Jin, Y. Lu, J. Qin, Z. Cheng, Z. Mao, Swiftids: real-time intrusion detection system based on lightgbm and parallel intrusion detection mechanism, Comput. Secur. 97 (2020) 101984.
[22] S. Kapp, J.-K. Choi, T. Hong, Predicting industrial building energy consumption with statistical and machine-learning models informed by physical system parameters, Renew. Sustain. Energy Rev. 172 (2023) 113045.
[23] S. Karatasou, M. Santamouris, V. Geros, Modeling and predicting building's energy use with artificial neural networks: methods and results, Energy Build. 38 (2006) 949–958.
[24] 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, Adv. Neural Inf. Process. Syst. 30 (2017).
[25] J.K. Kissock, C. Eger, Measuring industrial energy savings, Appl. Energy 85 (2008) 347–361.
[26] K. Li, B. Wei, Q. Tang, Y. Liu, A data-efficient building electricity load forecasting method based on maximum mean discrepancy and improved tradaboost algorithm, Energies 15 (2022) 8780.
[27] Q. Li, Q. Meng, J. Cai, H. Yoshino, A. Mochida, Predicting hourly cooling load in the building: a comparison of support vector machine and different artificial neural networks, Energy Convers. Manag. 50 (2009) 90–96.
[28] L. Lundström, F. Wallin, Heat demand profiles of energy conservation measures in buildings and their impact on a district heating system, Appl. Energy 161 (2016) 290–299.
[29] S. Makridakis, E. Spiliotis, V. Assimakopoulos, M5 accuracy competition: results, findings, and conclusions, Int. J. Forecast. (2022).
[30] R.F. Mustapa, N.Y. Dahlan, A.I.M. Yassin, A.H.M. Nordin, Quantification of energy savings from an awareness program using narx-ann in an educational building, Energy Build. 215 (2020) 109899.
[31] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, Front. Neurorobot. 7 (2013) 21.
[32] A.H. Neto, F.A.S. Fiorelli, Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption, Energy Build. 40 (2008) 2169–2176.
[33] N. Oses, A. Legarretaetxebarria, M. Quartulli, I. García, M. Serrano, Uncertainty reduction in measuring and verification of energy savings by statistical learning in manufacturing environments, Int. J. Interact. Des. Manuf. 10 (2016) 291–299.
[34] F. Petropoulos, D. Apiletti, V. Assimakopoulos, M.Z. Babai, D.K. Barrow, S. Ben Taieb, C. Bergmeir, R.J. Bessa, J. Bijak, J.E. Boylan, J. Browell, C. Carnevale, J.L. Castle, P. Cirillo, M.P. Clements, C. Cordeiro, F.L. Cyrino Oliveira, S. De Baets, A. Dokumentov, J. Ellison, P. Fiszeder, P.H. Franses, D.T. Frazier, M. Gilliland, M.S. Gönül, P. Goodwin, L. Grossi, Y. Grushka-Cockayne, M. Guidolin, M. Guidolin, U. Gunter, X. Guo, R. Guseo, N. Harvey, D.F. Hendry, R. Hollyman, T. Januschowski, J. Jeon, V.R.R. Jose, Y. Kang, A.B. Koehler, S. Kolassa, N. Kourentzes, S. Leva, F. Li, K. Litsiou, S. Makridakis, G.M. Martin, A.B. Martinez, S. Meeran, T. Modis, K. Nikolopoulos, D. Önkal, A. Paccagnini, A. Panagiotelis, I. Panapakidis, J.M. Pavía, M. Pedio, D.J. Pedregal, P. Pinson, P. Ramos, D.E. Rapach, J.J. Reade, B. Rostami-Tabar, M. Rubaszek, G. Sermpinis, H.L. Shang, E. Spiliotis, A.A. Syntetos, P.D. Talagala, T.S. Talagala, L. Tashman, D. Thomakos, T. Thorarinsdottir, E. Todini, J.R. Trapero Arenas, X. Wang, R.L. Winkler, A. Yusupova, F. Ziel, Forecasting: theory and practice, Int. J. Forecast. 38 (2022) 705–871.
[35] A.M. Prasad, L.R. Iverson, A. Liaw, Newer classification and regression tree techniques: bagging and random forests for ecological prediction, Ecosystems 9 (2006) 181–199.
[36] P. Price, Methods for analyzing electric load shape and its variability, Technical Report Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), 2010.
[37] F. Rossi, D. Velázquez, A methodology for energy savings verification in industry with application for a chp (combined heat and power) plant, Energy 89 (2015) 528–544.
[38] E. Sarmas, E. Spiliotis, N. Dimitropoulos, V. Marinakis, H. Doukas, Estimating the energy savings of energy efficiency actions with ensemble machine learning models, Appl. Sci. 13 (2023) 2749.
[39] E. Sarmas, E. Spiliotis, V. Marinakis, T. Koutselis, H. Doukas, A meta-learning classification model for supporting decisions on energy efficiency investments, Energy Build. 258 (2022) 111836.
[40] E. Sarmas, E. Spiliotis, E. Stamatopoulos, V. Marinakis, H. Doukas, Short-term photovoltaic power forecasting using meta-learning and numerical weather prediction independent long short-term memory models, Renew. Energy 216 (2023) 118997.
[41] R.E. Schapire, The boosting approach to machine learning: an overview, in: Nonlinear Estimation and Classification, 2003, pp. 149–171.
[42] C. Strobl, J. Malley, G. Tutz, An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests, Psychol. Methods 14 (2009) 323.
[43] S. Touzani, J. Granderson, S. Fernandes, Gradient boosting machine for modeling the energy consumption of commercial buildings, Energy Build. 158 (2018) 1533–1543.
[44] V. Tsiatsis, S. Karnouskos, J. Holler, D. Boyle, C. Mulligan, Internet of Things: Technologies and Applications for a New Age of Intelligence, Academic Press, 2018.
[45] D. Ürge-Vorsatz, L. Danny Harvey, S. Mirasgedis, M.D. Levine, Mitigating co2 emissions from energy use in the world's buildings, Build. Res. Inf. 35 (2007) 379–398.
[46] C. Wade, Hands-on Gradient Boosting with XGBoost and Scikit-Learn: Perform Accessible Machine Learning and Extreme Gradient Boosting with Python, Packt Publishing Ltd., 2020.
[47] Z. Wen, J. Shi, B. He, J. Chen, K. Ramamohanarao, Q. Li, Exploiting gpus for efficient gradient boosting decision tree training, IEEE Trans. Parallel Distrib. Syst. 30 (2019) 2706–2717.
[48] S.L. Wong, K.K. Wan, T.N. Lam, Artificial neural networks for energy analysis of office buildings with daylighting, Appl. Energy 87 (2010) 551–557.
[49] K. Ye, G. Demirezen, A. Fung, E. Janssen, The use of artificial neural networks (ann) in the prediction of energy consumption of air-source heat pump in retrofit residential housing, in: IOP Conference Series: Earth and Environmental Science, in: IOP Publishing, vol. 463, 2020, 012165.