Discover Internet of Things

Research

Machine learning model for predicting long-term energy consumption in buildings

Aseel Hussien[1] · Aref Maksoud[1] · Aisha Al-Dahhan[2] · Ahmed Abdeen[3] · Thar Baker[4]

Received: 26 October 2024 / Accepted: 24 February 2025
Published online: 03 March 2025

Abstract
The rapid growth in the construction sector has led to increased energy consumption and carbon emissions. Calculating energy usage and emissions is essential to energy security and promoting sustainable sector development. Therefore, the study objective is to investigate the utilazation of machine learning algorithm to predict long-term energy consumption in buildings sector, aiming to improve sustainable design and energy optimization, via the implementation of three machine learning models, XGBoost, Support Vector Regression, and Long-Short-Term Memory networks, to predict energy consumption. These models are adept at capturing complex interactions between building characteristics, environmental factors, and energy patterns. Although previous studies have explored various machine learning techniques for energy efficiency, limited research links these models to practical applications in building performance simulation. Furthermore, there is a lack of comparative evaluation of advanced machine learning models such as XGBoost, Support Vector Regression, and Long-Short-Term Memory to predict the energy consumption of building envelopes, particularly in hot climates such as the UAE. This research aims to fill this gap by providing a detailed comparison of these models against alternative approaches mentioned in the literature. The findings position Long-Short-Term Memory as a transformative force in predictive modeling, demonstrating exceptional precision with an R-squared value of 0.993 and a Mean Squared Error of 0.004. In contrast, Support Vector Regression and XGBoost showed limited predictive capabilities, with R-squared values of 0.462 and 0.94, respectively. This study establishes a solid data-driven foundation for architects and engineers to inform decisions on energy-efficient building designs, advocating Long-Short-Term Memory as the superior model for predicting energy performance.

Highlights

- Machine learning models enhance energy predictions, aiding sustainable building design and decision-making.
- LSTM outperforms other models in forecasting energy efficiency of wall materials in hot climates.
- Practical insights link ML theory to real-world applications, improving construction sustainability practices.

Keywords Machine learning · Energy consumption · Building envelop · Indoor environment quality

✉ Aseel Hussien, ahussien@sharjah.ac.ae; Aref Maksoud, amaksoud@sharjah.ac.ae; Aisha Al-Dahhan, u18103760@sharjah.ac.ae; Ahmed Abdeen, ahmed_saleem@uow.edu.au; Thar Baker, t.shamsa@brighton.ac.uk | [1]Department of Architectural Engineering, University of Sharjah, Sharjah, UAE. [2]Department of Computer Science, University of Sharjah, Sharjah, UAE. [3]Civil, Mining, Environmental, and Architectural Engineering, University of Wollongong, Wollongong, Australia. [4]School of Architecture, Technology and Engineering, University of Brighton, Brighton, UK.

1 Introduction

Energy conservation in the construction sector has become a critical area of focus due to the sector's significant energy consumption worldwide [1, 2]. Construction processes and building operations are responsible for approximately one-third of global greenhouse gas (GHG) emissions, making the sector a substantial contributor to global warming [3, 4]. This stresses the urgent need to adopt energy-efficient strategies and sustainable building operations, including design and materials. These strategies enhance energy efficiency and thermal comfort, reduce overall energy consumption, and significantly lower CO_2 emissions [5–7]. Several solutions, including machine learning (ML) algorithms, have been proposed to address these challenges. ML is a process through which machines can independently learn, adapt, and improve their performance based on existing data without explicit programming. Therefore, ML can manage vast amounts of data, learn from historical energy consumption patterns, and predict future energy needs more accurately. This adaptability enables more efficient building design by accounting for various variables, including construction materials, construction style, and building envelop, which are often difficult to predict using conventional methods. Therefore, ML can predict energy usage across various building types from the early stages before being constructed [8]. This adaptability of ML has led to growing research interest in the construction sector, where ML is now being explored for multiple applications, including project management, site safety, risk mitigation, cost estimation, and schedule management [9]. In regard to energy efficiency, ML is particularly promising, as it enables the analysis of building performance data to optimize energy consumption, predict maintenance needs, and enhance overall building operational efficiency [10]. By utilizing ML into energy management strategies, the construction sector can make significant improvements toward reducing its environmental impact while improving the sustainability and functionality of the built environment. Thus, the rise of ML and greater data accessibility in buildings present outstanding opportunities for its application in building energy system modeling and analysis [11]. Several studies have attempted to investigate the prediction via ML models within the construction sector, including different models like Gradient Boosting (GB), Gradient Boosted Regression Trees (GBRT), Random Forest (RF), XGBoost, among others, to predict resource management, safety risk assessments, and construction cost estimates (Table 1).

However, research revolving around energy consumption caused by buildings seems limited regardless. Our previously published study [12], utilized random forest (RF) and multi-liner recognition (MLR) models to predict the energy performance of building envelopes, this work explores the potential of rigorous ML techniques to accurately predict the energy consumption of building envelopes. Recognizing the importance of precise energy consumption predictions to optimize building performance and promote environmental sustainability. Hence this research aims to advance these capabilities to introduce and evaluate the efficacy of different ML algorithms in this domain, including XGBoost, Long-Short-Term Memory Networks (LSTM), and Support Vector Regression (SVR). The selected ML models are chosen as they cover a wide range of scenarios predicting energy consumption [13, 28]. They outperform other models by utilizing their unique strengths in feature interaction modeling, handling non-linearities, and capturing time-based dependencies [13, 28]. Table 1 compares the ML models selected (XGBoost, LSTM, SVR), including strengths, performance, and usage for predicting building envelope energy consumption.

The novelty of XGBoost, SVR, and LSTM models lies in their adapted capabilities to tackle diverse challenges in energy consumption prediction, from handling non-linear relationships (XGBoost) and noisy datasets (SVR) to modeling time-based dependencies (LSTM). They align with existing research gaps by offering a more accurate, interpretable, and all-inclusive approach to predicting building envelope energy consumption, enabling advancements in energy efficiency, sustainability, and energy management strategies. As such, utilizing multiple ML models (XGBoost, SVR, and LSTM) provides a broader perspective on their performance, provides a more comprehensive view of the model's effectiveness, and enhances decision-making processes in construction and sustainability efforts. Furthermore, the study used the Mean Squared Error (MSE) and pseudo-R-squared metrics to assess model performance. For a fair comparison with the MLR and RF models of our previous paper, all models were tested using encoded data that underwent one-hot encoding, as there is no inherent order/ranking to the categorical data, using the same dataset from our previous research study in [12] of wall materials for building envelopes with a total of 1119 data of different wall options. Each stage was carefully designed to understand the significance of each variable on the building's overall energy efficiency. the study then conducted a grid search and cross-validation to fine-tune the parameters of each model, optimizing their performance.

This study, therefore, aims to utilize ML algorithm, including XGBoost, SVR, and LSTM to forecast the energy performance of external wall materials over long periods within the hot climate of the UAE. Consequently, the study contributions involve:

Table 1 the strength, performance, and usage of XGBoost, SVR, and LSTM for Predicting Building Envelope Energy Consumption

ML Model	Strength	Performance	Usage	References
XGBoost	Handles non-linear relationships effectively. Provides feature importance for interpretability. Scalable and computationally efficient for large datasets.	Excels on structured data. Captures interactions between features. Robust against overfitting with regularization.	Ideal for modeling static or slow-changing factors (e.g., material properties). Provides insights into key drivers of energy consumption.	[13, 23, 26]
SVR	Excels in modeling complex, non-linear relationships. Performs well with small and messy datasets. Built-in regularization prevents overfitting	Handles detailed and intricate relationships.	Effective when energy consumption is influenced by subtle and intricate factors. • Performs well without requiring extensive data.	[18, 22]
LSTM	Specializes in sequential and time-series data. Captures long-term dependencies and patterns handles non-linear and dynamic systems efficiently.	Outperforms traditional models for time-dependent predictions (e.g., daily/seasonal energy use trends).	Best suited for capturing temporal dependencies in energy consumption (e.g., patterns over time-of-day, weather, or operational schedules). Offers unmatched forecasting accuracy for sequential data.	[14, 28, 29]

- Presenting a comprehensive review of different ML models used in existing research to pave the way for future work on this study.
- Applying three distinct ML models including XGBoost, SVR, and LSTM in a unified framework represents a holistic approach to solving the problem of energy consumption prediction.
- Bridging a critical gap in the literature by linking theoretical ML models with practical applications in building performance simulation, providing actionable insights that can be immediately applied in real-world scenarios, and enhancing the practical value of academic research in building technology.
- Conducting a comparative analysis of the performance of XGBoost, SVR, and LSTM models against RF, MLR, and other models reported in the literature. This helps to identify the strengths and limitations of each approach. Providing a solid, data-driven foundation for architects and engineers to make informed decisions. The study advocates LSTM as the superior model for predicting energy performance.

The paper is organized as follows: Section 2 presents a literature review, and the detailed proposed methodology is presented in Section 3. Section 4 demonstrates the results and discusses the statistical outcomes, including the performance evaluation of three ML models (XGBoost, SVR, and LSTM) used to predict energy consumption. Finally, in Section 5 the research conclusion is presented.

2 Literature review

Many researchers investigated various ML techniques, including SVM, ANN, GLM, DNN, RF, and GB (Table 2). The Comparative analysis in [13, 14] shows that GB surpasses other machine learning algorithms in terms of accuracy and performance, the importance of energy-efficient buildings and the challenges associated with estimating their energy consumption during the initial phases of project development. While a study by [15] evaluates the performance of five ML models in predicting heating and cooling loads in buildings, comparing them with two simulation tools: EnergyPlus and Ecotect. The models are ANN, SVM, GP, RF, and GBRT, and employ grid-search with cross-validation for parameter optimization and sensitivity analysis for important input variables. Their results showed that GBRTs provide the most accurate predictions, while SVMs are preferable for more straightforward data sets due to their simplicity and speed. The researcher in [16] presents an integrated model for designing energy-efficient building envelopes. The model consists of three main components: a simulation model, a predictive model utilizing ML algorithms like RF, ANN, DNN, SVM, GENLIN, and GB (with GB identified as the most suitable), and an optimization model employing AI optimization algorithms. The model achieved significant cost and energy savings. The study in [17] explores data-driven methods for building envelope design to improve energy efficiency. It quantifies envelope impacts on heating and cooling energy usage using feature selection and game-theoretic approaches. RF classifiers aid in analysis, and the results showed energy savings of 1.05%–21.2% compared to baseline.

Several studies have pointed towards the effectiveness of the XGBoost model in predicting building energy consumption through attributes such as their envelope. The authors of [23] propose a ML framework to manage energy in buildings with thermally anisotropic envelopes, addressing computational cost concerns. It trains models like XGBoost, MLR, and DNN to predict heat fluxes. The results demonstrate XGBoost's efficiency, achieving high accuracy comparable to DNN with significantly less training time. While the research study in [13] develops a heating demand prediction model using RF, XGBoost, and ET algorithms. When XGBoost is the most accurate, with an average mean absolute percentage error (MAPE) of around 40%, it identifies crucial input variables and demonstrates the potential to integrate into renovation programs to enhance building energy efficiency. The researchers in [26] explored using ML models, including SVR, RF, and XGBoost, to predict building energy consumption efficiently and accurately. The results indicate that XGBoost showed superior performance in regression tasks, offering a promising solution to reduce computational costs and providing valuable insight into optimizing building design strategies.

While SVR and SVM have shown superior performance in predicting complex patterns in building energy consumption compared to other models. The authors in [14] proposed a model to predict energy consumption in residential buildings integrated with PCM in the Mediterranean climate using MR, SVM, and ANN. Sensitivity analysis identifies key parameters such as PCM melting point and shape factor. Statistically, SVM and ANN show reliability with R-squared values over 0.99. The key influencing parameters include cooling/heating degree days, volume, and PCM melting point. The study in [19] introduces a novel vector field-based SVR method to improve accuracy, robustness, and generalization in building energy consumption prediction, optimizing the feature space to approximate non-linearity with linearity,

Table 2 The different ML algorithms used within the construction sector

ML model implemented in the construction sector	Applications in construction sector	Approximate accuracy	References
Support vector machine (SVM)	Defect detection, predicting material properties, optimizing resource allocation	75–90%	[18–22]
Artificial neural networks (ANN)	Predicting construction costs, structural health monitoring, project scheduling, and energy consumption forecasting	80–95%	[14–16, 22, 24]
Deep neural networks (DNN)	Image recognition for site monitoring, quality control, predictive maintenance	85–98%	[16, 23]
Generalized linear models (GLM)	Modeling relationships in construction data, predicting project delays or cost overruns	70–85%	[25]
Random forest (RF)	Risk assessment, prediction of material strength, and detection of structural anomalies	80–95%	[13, 15, 16, 26]
Gradient boosting (GB)	Predictive analytics, cost forecasting, scheduling, risk analysis	80–95%	[15, 16, 19]
Gradient boosted regression trees (GBRT)	Precise predictions in cost estimation, resource management, and safety risk assessments	85–95%	[15, 22]
Generalized linear mixed models (GENLIN)	Handling hierarchical data structures in construction projects	70–85%	[16]
XGBoost,	High-performance predictive modeling, predicting project outcomes and costs	85–97%	[13, 23, 26]
Multiple linear regression (MLR)	Cost estimation, project time prediction, and analysis of factors that affect construction processes	65–80%	[23]
Extra trees (ET)	Feature Selection, Predicting Material Properties, Modeling Project Timelines	80–95%	[13]
Multiple regression (MR)	Predicting construction outcomes based on multiple variables.	65–80%	[14]
Linear regression (LR)	Estimation of costs and prediction of timelines based on historical data	60–75%	[27]
Long short-term memory (LSTM)	Time series predictions, forecasting construction demand, equipment maintenance schedules	85–98%	[14, 28, 29]
Deep learning (DL)	Image recognition, predicting outcomes, automated design assessments	85–98%	[28, 29]
Sequence-to-sequence (Seq2Seq)	Sequence prediction, project scheduling, and resource planning	80–95%	[24]
WaveNet (WN)	Time series predictions, forecasting labor requirements, equipment usage	85–98%	[15, 20]
Random tree (RT)	Classification, regression tasks, prediction of construction defects, project risks	75–90%	[22, 27]
Deep reinforcement learning (DRL)	Optimizing Planning, Autonomous Equipment Operation, and Decision-making in Dynamic Environments	80–95%	[28]
Support vector regression (SVR)	building energy forecasting	80% to 95%	[18, 22]

this method outperforms traditional approaches when applied to summer hourly cooling load data from a large office building in a coastal town of China. The researchers in [27] investigated improving building energy efficiency to reduce CO2 emissions using an open-source data set, and the study used investigative data analysis and feature engineering to refine data for model training. They used LR and SVR with repeated cross-validation. The SVR model outperformed the LR model with a higher accuracy of 92.41% and a lower MAPE of 7.59%. The SVR model also surpassed a more complex model from a previous study, proving it more effective in predicting building energy consumption.

Similarly, LSTM models have also demonstrated their effectiveness in predicting building energy consumption, offering high accuracy in handling time-series data related to energy use patterns. The study in [16] focuses on enhancing short-term building energy prediction using LSTM-based models. It evaluates pure LSTM and improved versions (LSTM-CNN, CNN-LSTM, LSTM-Attention, CNN-Attention-LSTM, LSTM-Attention-CNN) with data from Building Data Genome Project 2, testing on 60 buildings from various climates and types. The results revealed that parameter optimization reduces RMSE by 6.2% to 29.2%. CNN-LSTM reduces RMSE by 2.9% with one-year data, while LSTM-Attention provides the most stable performance with two-year data, reducing RMSE by up to 5.6%. The study [29] developed a deep learning interval forecasting model for building energy consumption using fuzzy information granulation, attention mechanism, and LSTM. The model, validated with real data, shows that the size of the granulation window affects the prediction intervals. Attention-based LSTM outperforms conventional LSTM in prediction accuracy and interval coverage. In addition, the study in [24] improved the prediction of building energy demand using detailed data and the EnergyPlus simulation program, incorporating actual weather and occupancy data. Validation with IoT-based data reduced the mean bias error to 4.73% and Cv(RMSE) to 12.26%. A sequence-to-sequence model with LSTM cells achieved an RMSE of 4.48% and a weighted average error of 3.07% with occupancy information and solar radiation being the most influential factors in energy demand prediction.

With the various machine learning models implemented within the construction sector, there is a clear focus on some models with less focus on other ML models within the sector.

Figure 1 illustrates the number of publications focused on different ML models used within the construction sector extracted from the Scopus database. This data highlights key trends and preferences in the application of ML in construction, which are increasingly leveraging these models to improve efficiency, safety, and decision-making.

Key observations from the data showed that the dominance of Random Forest, with 331 publications, reflects its wide acceptance in the sector, which is ideal for construction applications such as risk assessment, predictive maintenance, and quality control. Further, Fig. 1 showed a significant focus on deep learning models with 269 publications, reflecting the surge in interest and applications of deep learning techniques across various domains, indicating a strong trend toward using advanced neural networks for complex tasks in the sector. These models excel at analyzing unstructured data like images and videos, which are common in construction site monitoring, defect detection, and progress tracking. Figure 1 shows the popularity of XGBoost with 113 publications, is another widely researched model due to its efficiency, speed, and performance in structured tabular data. In addition, XGBoost is widely used for predictive modeling tasks in construction, such as energy consumption, cost estimation, schedule forecasting, and resource optimization. While moderate use SVM and ANN also feature prominently, with 74 and 132 publications, respectively. These models have been fundamental in ML research, providing strong baselines for classification and regression tasks. Both maintain relevance in specific applications such as the classification of construction materials, the classification of safety risks, and the regression tasks to predict structural performance.

Additionally, an emerging trend in advanced models LSTM with 42 publications and deep reinforcement learning with 18 publications points towards emerging trends in sequence modeling and decision-making tasks, such as time series forecasting, which is critical in project scheduling and monitoring. In contrast, reinforcement learning is explored to optimize construction robotics and automation. However, the limited use of traditional and specialized models in GLM, GENLIN, GBRT, and Seq2seq with fewer publications indicates their niche application or declining popularity compared to more flexible and powerful models like Random Forest and deep learning variants.

The data in Fig. 1 reflect the strategic adoption of ML models in the construction sector to address key industry challenges. The dominance of ensemble methods like RF and XGBoost and the growing use of DL models indicate a strong focus on predictive accuracy, automation, and handling complex unstructured data. This trend drives significant improvements in risk mitigation, resource optimization, and project management safety, aligning with the sector's goals of improving productivity and reducing costs. The publication trends suggest that the construction industry is adopting ML models and actively exploring and integrating these technologies into their workflows, signaling a transformative shift towards data driven construction management and operational excellence.

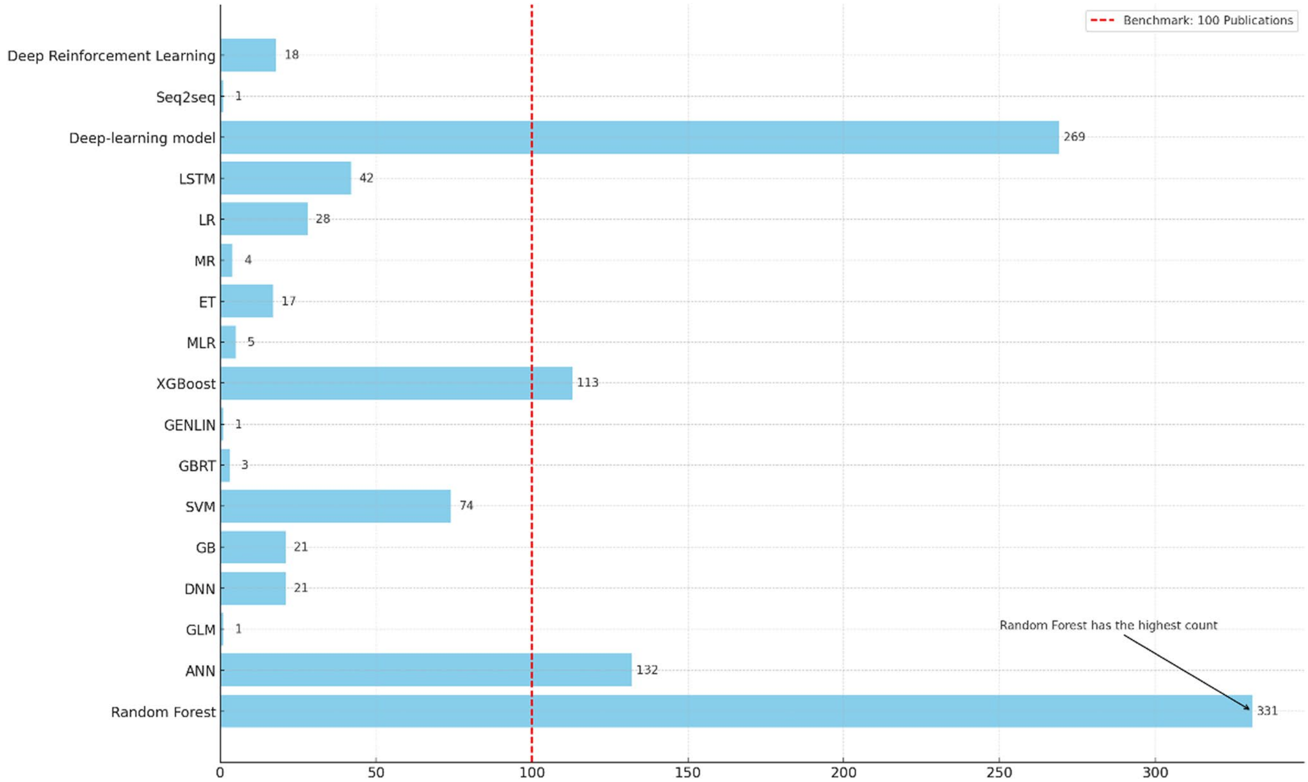


Figure 1 The number of publications for various Machine Learning (ML) models, highlighting the popularity and research focus on specific algorithms within the field (created by the author)

However, Fig. 2 shows that the adoption of ML in construction is lower than in industries driven by technology, with 10%. Current applications include predictive maintenance of equipment, safety monitoring, project scheduling, and cost estimation. The relatively slower adoption rate can be attributed to challenges such as inconsistent data, slower digital transformation, and the sector's traditionally conservative approach to adopting new technologies. While ML is being used significantly in the construction industry, there are notable limitations and gaps in research and application compared to more advanced sectors such as finance, healthcare, and manufacturing. Figure 2 highlights the different paces of ML adoption in different sectors, with technology-driven industries leading the way.

Our previous research [12] explored the various attributes of wall materials used in building envelopes, particularly in the hot climate of the UAE, with a primary focus on their impact on indoor thermal comfort and efficient energy use. The objective was to employ ML techniques to forecast the energy performance of various wall materials of the building envelope over a long period of time. Building on the insights presented in the previous study, this research embarks on the potential of ML techniques to efficiently predict the energy consumption of building envelope. Recognizing the significance of accurate energy consumption predictions in optimizing building performance and environmental sustainability.

3 Methodology

This paper details how ML models, including XGBoost, SVR, and LSTM, are trained using data from building envelopes simulated in two stages:

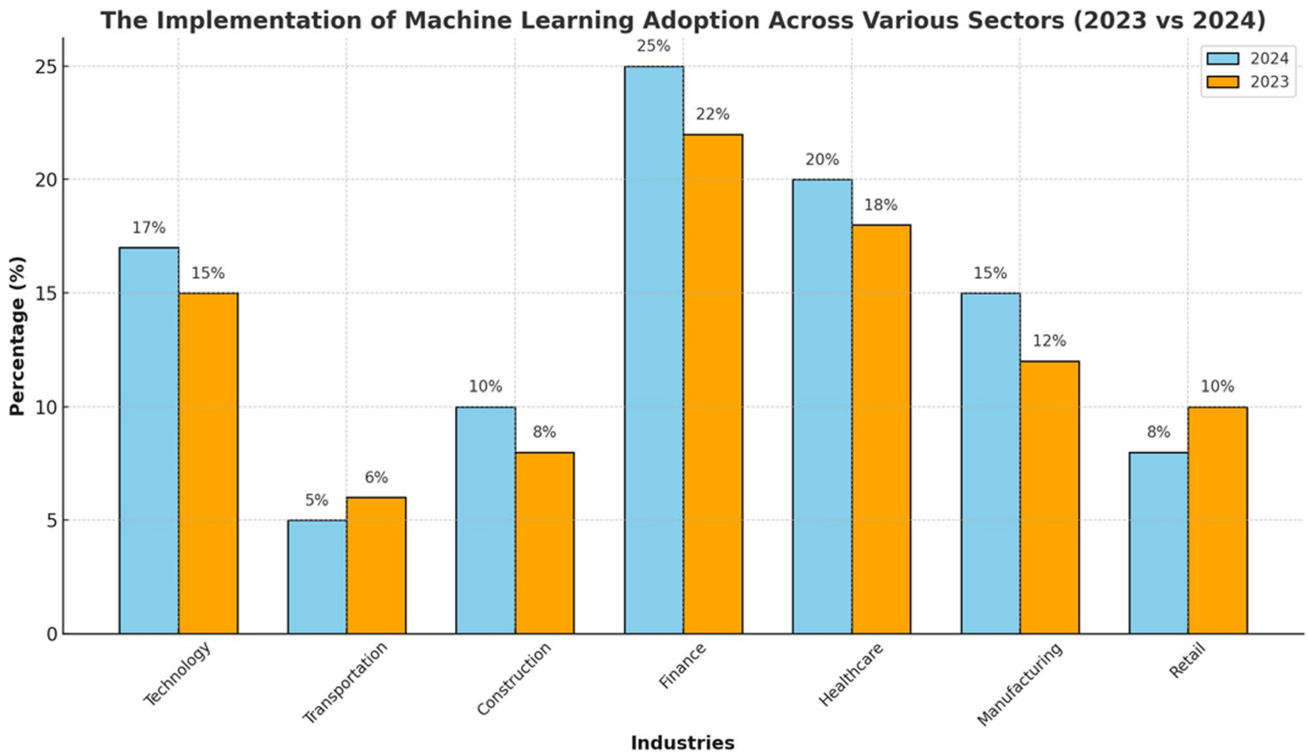


Fig. 2 Comparison of Machine Learning Adoption Rates Across Various Industries: 2023 vs. 2024 (created by author)

3.1 Phase 1 data generation

Involves generated data through building performance analysis using IES-VE software in various wall construction scenarios, with the results saved as a CSV file for the subsequent stage. These data were extracted from previous work [12]. The simulation consisted of several scenarios, presenting the various options explored for the building envelope, each designed to assess different variables affecting building performance. The various options are listed below:

- Different Wall Materials
- Different Thicknesses of the Wall
- Different Orientation of the Space
- Different Wall Construction
- Different Shape Factors

Wall materials and thickness impact insulation and heat transfer, while orientation and shape affect solar exposure and heat regulation. Thermal transmittance and resistance values indicate the ability of walls to retain or lose heat, influencing heating and cooling needs. The thermal mass regulates temperature fluctuations, helping maintain comfort and reduce energy consumption. These factors collectively play a key role in designing energy-efficient, sustainable buildings.

Table 3 shows a sample of the data used in the simulation when wall thickness, orientation, material, shape, thermal transmittance (U-value), thermal resistance (R-value), thermal mass, and yearly energy consumption are crucial for determining the energy efficiency and thermal performance of building envelopes.

For example, a wall built with HCB was simulated with 200 mm thickness and 0 orientation (North), with wall construction W1 and shape A. then HCB was simulated with 200 mm thickness and 0 orientation (North), with wall construction W1 and shape B. A total of 1119 data were stimulated according to the different options for each wall.

Table 3 A sample of walls details used in the simulation with different Thermal Performance and energy consumption [12]

Wall type	Wall thickness (mm)	Orientation	Material	Shape	Thermal transmittance (U-value) m_2K/W	Thermal resistance (R-value)W/m^2K	Decrement factor/ summer Day	Thermal mass kj/m2k	Yearly energy consumption MWh
W-1	222	N	ACB	A	1.64	0.43	0.33	136.89	11.74
W-2	222	N	ACB	A	0.91	0.53	0.13	8.53	10.72
W-3	222	N	ACB	A	0.96	0.72	0.17	156.80	9.72
W-4	222	N	ACB	A	1.83	0.96	0.52	154.74	12.25
W-5	222	N	ACB	A	1.22	0.83	0.63	202.10	9.52
W-6	222	N	ACB	A	0.93	0.98	0.42	174.92	8.34
W-7	222	N	ACB	A	0.63	1.26	0.53	90.63	8.62
W-8	222	N	ACB	A	0.74	1.96	0.64	264.95	11.71
W-9	222	N	ACB	A	0.61	1.48	0.52	207.92	10.00
W-10	222	N	ACB	A	0.64	1.38	0.17	73.05	10.31
W-11	222	N	ACB	A	0.84	1.31	0.09	72.77	9.26
W-12	222	N	ACB	A	0.71	1.23	0.21	73.05	9.62
W-13	222	N	ACB	A	0.63	0.88	0.12	78.23	9.81

3.2 Phase 2: ML algorithm utilization

In this phase (Fig. 3), the dataset was divided into training (70%) and testing (30%) samples; this ratio is used because it offers a good balance between training the model and evaluating its performance on unseen data (the unseen data here refers to the data that the model has not encountered during training, to help in simulating real-world inputs ensuring the model can generalize beyond the patterns in the training set to allows for an unbiased evaluation of how well the model will perform on real-world data), using tenfold cross-validation [30]. Model bias can also stem from the algorithms themselves. For example, XGBoost may favor features with higher variance, which can lead to biased predictions. Additionally, preprocessing steps like one-hot encoding may introduce biases in models such as SVR and LSTM. To mitigate these biases, we ensured that the data is diverse and that preprocessing techniques do not negatively impact the model's ability to generalize to unseen data. Furthermore, complex models like XGBoost are more prone to overfitting, especially when trained on smaller datasets. This overfitting can be reduced through techniques such as cross-validation and regularization, which help the model better generalize to new data.

Additionally, widespread simulations are conducted to fine-tune the model parameters, enhancing the prediction of building energy efficiency using machine learning algorithms.

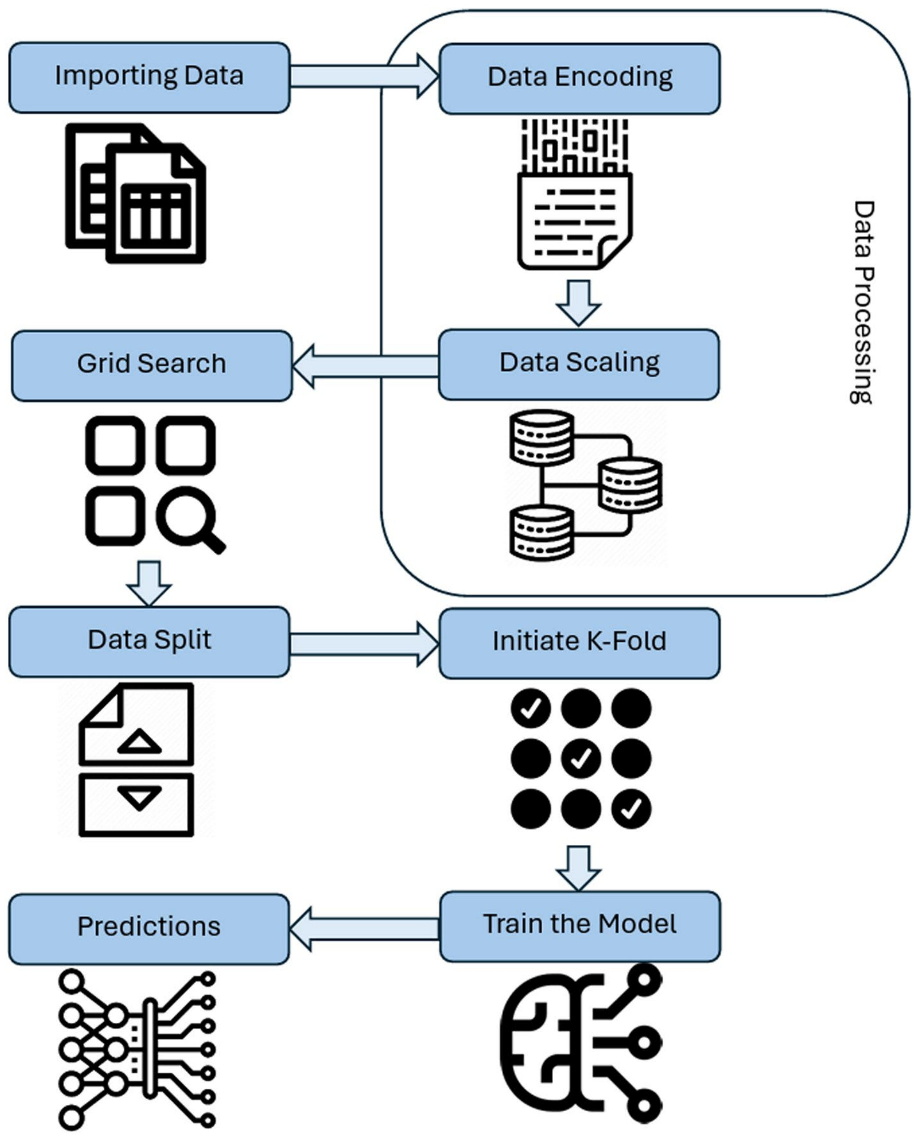


Figure 3 Overview of the ML stage of the research methodology

In this phase, three ML models, XGBOOST and SVR, and LSTM, are used to predict building envelope energy consumption. These models can capture complex relationships between various characteristics of buildings, environmental factors, and energy consumption patterns.

As shown in Fig. 3, the process starts by importing the data into a software environment where they can be analyzed. These data will be the same as our previous study's [12]. The data are processed with one-hot encoding to convert categorical variables into numerical format. This involves creating binary columns for each unique category present in the categorical variables 'Construction Type,' 'Material' and 'Shape.' The encoded variables can be effectively used to compare the ML algorithms being used in this study with the ones used in the previous study where we tested MLR and RF, MLR being a model that cannot process categorical variables unless they underwent one-hot encoding.

After that, the data is scaled, where the range of values of variables is adjusted to a standard scale between 0 and 1. It should be noted that scaling was only applied when the data was used in SVR and LSTM, as SVR and LSTM rely on distance-based measurements, where the magnitude of the input features can significantly affect the model's performance. In contrast, XGBoost is less sensitive to feature scaling due to its tree-based structure. Additionally, SVR measures how far apart data points are, and any slight change of the same value in any numeric feature is given the same importance, making it essential to scale the data for optimal performance. LSTM is a model that uses sigmoid and tanh functions sensitive to magnitude changes, so variable values must be scaled.

After scaling, the next step involves performing a grid search to find the optimal hyperparameters for the models, by selecting the ranges for each model's hyperparameters based on previous studies and empirical performance. Grid search methodically examines a specified range of hyperparameter values to determine the combination that delivers optimal performance for each model. For XGBoost, learning rates from 0.01 to 0.2 were tested, including the learning rate, maximum depth, number of trees, subsample, and column subsample. while for SVR, we considered kernel types such as RBF and polynomial with values for the regularization parameter (C) between 1 and 100.

For LSTM they were the number of LSTM units, dropout rate, activation function, learning rate, number of epochs and batch size.

The data will be split into training and testing sets after hyperparameter tuning. Subsequently, KFold cross-validation will be initiated. Kfold divides the dataset into K subsets, trains the model K times, and uses a different fold as the validation set in each iteration. This approach helps to assess the robustness and generalization performance of the model. During each iteration, the model will be trained in a distinct training set and evaluated in a separate validation set. This process ensures a comprehensive evaluation of the model's performance in various data partitions, contributing to a more reliable assessment of its effectiveness and generalizability.

The models will lastly be evaluated" based on their performance metrics. Two common metrics used for evaluation are MSE and pseudo R-squared. MSE measures the average square difference between the actual and predicted values, providing information on the model's predictive accuracy. In addition to the performance metrics, we report error margins to understand better the variability in model performance across different folds and subsets of data. For instance, the pseudo R-squared value for XGBoost in the testing set was 0.91 with an error margin of ± 0.03, indicating a moderate level of variability in its ability to explain the variance in energy consumption. Similarly, the MSE for the SVR model was 0.5 with a margin of error of ± 0.07. By incorporating these error margins, we can better appreciate the stability and consistency of model performance. These margins also help communicate the reliability of the results, providing a more detailed perspective on the model's accuracy and predictive power.

On the other hand, the pseudo R-squared assesses the proportion of variance explained by the model, indicating its overall goodness-of-fit. Each model will calculate these metrics to gauge its performance and determine the most effective approach to predict energy consumption in the building. Mathematically, MSE can be represented as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{1}$$

where n is number of data points, Y_i are the observed values, and $\hat{Y}_i$ are the predicted values.

3.2.1 Extreme gradient boosting (XGBOOST)

XGBoost, a robust and efficient model, is well-suited for energy consumption prediction tasks where the relationships between input features (e.g., building characteristics, environmental factors) and output target (energy consumption) are complex and non-linear. Since the RF model performed well in an earlier analysis by the author [12], using XGBoost is the next logical step, considering that it is also a decision tree-based model. XGBoost functions as an iterative decision tree algorithm that includes multiple decision trees, each tree improving the residuals of the trees that preceded it. Furthermore, XGBoost incorporates built-in regularization techniques to avoid overfitting, especially when working with complex datasets. Additionally, feature importance scores from the model can be used to perform feature selection, reducing the complexity of the model and enhancing its generalization ability. The predicted output of XGBoost is the sum of all individual tree results, as shown in Fig. 4.

The prediction average of the XGBoost model can be represented mathematically as:

$$\hat{y}_i = \sum_{k=1}^{n} f_k(x_i) \tag{2}$$

In this context, f_k corresponds to a specific tree, hence $f_k(x_i)$ represents the result of tree k, and $\hat{y}_i$ denotes the predicted value of the i the instance x_i.

XGBoost serves as a powerful and versatile ML algorithm optimized for accuracy and speed. This algorithm efficiently generates high-dimensional feature partitions through a strategic approach that builds on the errors of the preceding models. Similarly to RF, XGBoost utilizes an ensemble learning approach, employing multiple weak models, specifically decision trees. Each tree is trained on bootstrapped samples from the dataset, and through an aggregation process, the final prediction for each observation is computed as the average outcome across all models.

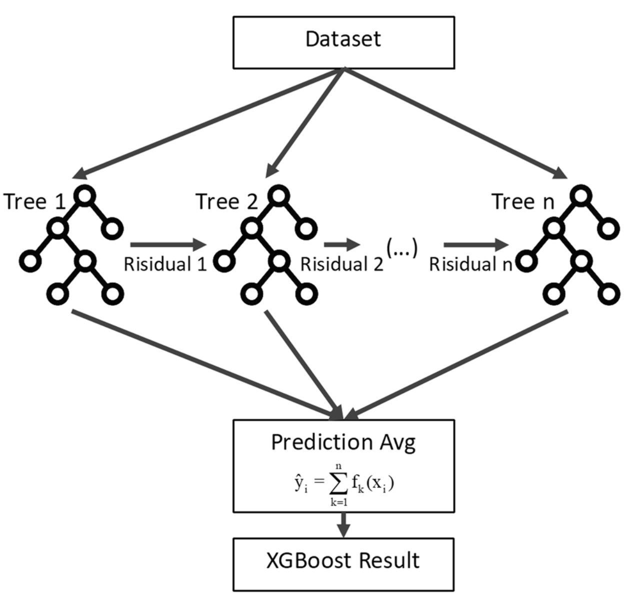
XGBoost employs boosting, which does not take a random subset of features to build decision trees like RF does, instead, it sequentially minimizes the errors of previous models while simultaneously boosting the influence of high-performing models. The XGBoost regressor starts with an initial prediction that is later modified based on the loss function to minimize the error. A good prediction would give a low value when computed using the loss function in Eq. 3.

$$\sum_{i=1}^{n} L(y_i, p_i) = \frac{1}{2}(y_i - p_i)^2 \tag{3}$$

where y_i are the observed values on the y-axis and p_i are the corresponding predictions. Through the concept of bagging, XGBoost creates an ensemble of multiple regressors with diverse structures. This diversity includes variations in the depths of the trees, the number of trees, and the subsets of input attributes. The aggregation of results from these diverse regressors contributes to a reduction in prediction error and enhances the overall robustness of the model.

An essential aspect of the XGBoost methodology involves the use of bootstrapping (sampling with replacement) to facilitate splits among different input attributes at varying levels, ensuring that decision trees remain highly uncorrelated. The algorithm's performance is quantitatively evaluated using metrics such as MSE, which quantifies the

Fig. 4 Architecture of the XGBoost model (created by author)



distance of each node from the actual value predicted by the model. Additional technical details can be explored in related work for deeper insights into decision tree structures and the feature bagging approach [31].

3.2.2 Long-short-term memory (LSTM)

The LSTM architecture comprises a memory unit consisting of four feedforward neural networks: forget gate, input gate, output gate, and candidate memory as shown in Fig. 5. These gates control the flow of information in memory operations, including deletion, insertion, and utilization. Input to the LSTM includes cell state (C) and hidden state (H) vectors from the previous time step, along with an external input vector (X). The LSTM regulates information flow through gating mechanisms to update cell and hidden states, serving as long-term and short-term memory. Gates uses sigmoid activation functions to create selector vectors for information retention.

Additionally, a candidate's memory generates candidate information that can be added to the cell state. The output gate determines the hidden state outputted by the LSTM, facilitating stability control over time. This architecture effectively manages information flow, enabling modeling both short-term and long-term memory-based behaviors.

LSTM is a powerful neural network architecture that is particularly effective in handling sequential data and finding widespread applications in tasks such as time-series analysis, classification, and regression. Its robust capabilities make it a cornerstone architecture of recurrent neural networks (RNNs), addressing challenges such as the vanishing gradient problem that often hinders traditional RNNs.

Unlike traditional RNNs, LSTM incorporates specialized memory cells and gating mechanisms. These additions enable the model to effectively combat the gradient's disappearance problem. The gates in LSTM are sigmoid activation functions that can only output a value between 0 and 1, this output determines whether a feature will be kept or discarded. There are 3 gates in LSTM: input gate Eq. 4, forget gate Eq. 5, and output gate Eq. 6.

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \tag{4}$$

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f) \tag{5}$$

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \tag{6}$$

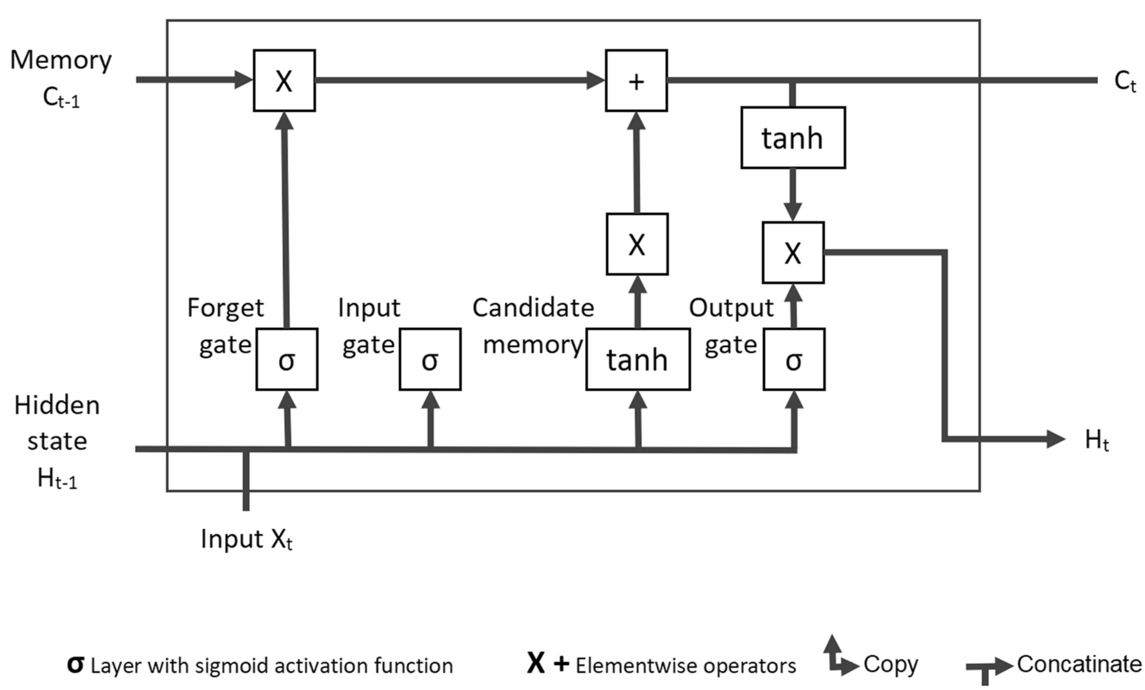


Fig. 5 Architecture of the LSTM model (created by author)

where σ is the sigmoid function, w is the weight of the respective gate, h_{t-1} is the output of the previous LSTM block, x_t is the input to the current block, and b is the bias for each gate. The LSTM model is composed of internal states that serve as memory cells, capable of retaining and updating information over long sequences. This unique feature enables LSTM to learn and memorize intricate long-range temporal patterns proficiently, making it a preferred choice for complex data sequences.

Within the ensemble learning framework, LSTM's adaptability shines through in its ability to handle varying depths, input sequences, and memory cell structures. This flexibility allows LSTM to capture optimal temporal dependencies, ensuring it remains a versatile solution for a wide array of sequential data tasks. Although energy consumption data does not inherently follow sequential patterns, using LSTM is justified as it helps capture temporal correlations between environmental factors, building configurations, and energy consumption over time, particularly when evaluating time-series data such as daily or seasonal energy usage.

The training process for LSTM involves carefully adjusting parameters to achieve optimal outcomes. Parameters such as the number of memory cells, the learning rate, and the sequence length play a pivotal role in fine-tuning the model. In sequential data analysis, LSTM's performance is often evaluated using metrics customized for such scenarios, including MSE. These metrics provide a quantitative assessment of the model's predictive accuracy and its ability to capture sequential patterns.

For more in-depth technical insights into the architecture and training process of LSTM, additional details can be explored in related works and literature on RNNs and sequence modeling.

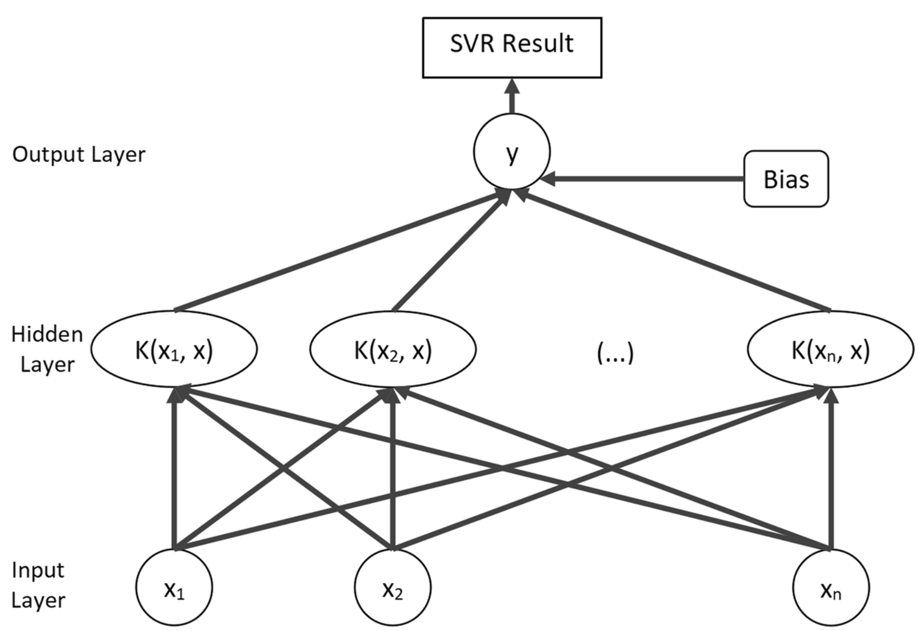
3.2.3 Support vector regression (SVR)

SVR is a versatile model that performs well in high-dimensional spaces. SVR can handle non-linear relationships in datasets using a wide range of kernel functions, such as polynomial, radial basis function (RBF), or sigmoid kernels, making it suitable for diverse energy prediction tasks. Figure 6 depicts the structure of the SVR model. A hidden layer connects the input and output layers, and its configuration is automatically determined based on the dataset. In this context, the kernel function is denoted by $K(x_i, x)$, it also represents the output of the i the hidden node for the input vector x. It signifies the mapping of input x and the support vector x_i through the selection of the kernel function.

SVR excels at separating data into distinct classes by strategically finding an optimal hyperplane that maximizes the margin between these classes. This algorithm performs exceptionally well in situations with complex decision boundaries and high-dimensional feature spaces.

At its core, SVR transforms input data into a higher-dimensional space to reveal an optimal hyperplane that effectively separates different classes. This hyperplane resembles a tube, where any point that falls within the tube will be

Fig. 6 Architecture of the SVR model (created by author)



disregarded as an error. The edges of this tube lie at a distance of epsilon; hence it's referred to as the ε-insensitive tube. The aim of SVR is to minimize the output of Eq. 7.

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{7}$$

where $\|w\|$ is the magnitude of the normal vector to the approximated surface, C is the regularization parameter, ξ_i and ξ_i^* are slack variables that fall above and below the tube respectfully where errors are regarded. This transformation is based on kernel functions, which play a crucial role in enabling SVR to capture intricate relationships and patterns in the data. The strength of SVR lies in its graceful handling of non-linearities, making it suitable for a variety of real-world scenarios where relationships between variables are not strictly linear.

The SVR's flexibility is noteworthy, shaped by key parameters influencing its behavior. The choice of kernel function, regularization parameters, and the specific characteristics of the chosen kernel contribute to tailoring the SVR to the nuances of the given problem. This flexibility places SVR as a versatile and robust tool in the ML landscape. Similarly to ensemble learning models, SVR benefits from careful parameter tuning to enhance its predictive abilities. Techniques such as grid search and cross-validation play crucial roles in this refinement process. The strategic adjustment of the parameters ensures that the SVR is effectively adapting to the intricacies of the data, achieving an optimal balance between bias and variance. In essence, the applicability of SVR spans various domains, making it a preferred choice for tasks were understanding complex relationships and achieving accurate predictions are crucial. To gain deeper technical insights into SVR, explore additional details in the literature on ML algorithms, particularly SVR and its applications [32]

In summary, the models differed in their approaches. XGBoost is similar to RF in that both utilize ensemble learning techniques based on decision trees. SVR, a supervised learning algorithm seeks to identify the optimal hyperplane that distinguishes between classes of data points. LSTM a form of RNN is specifically designed to recognize long term dependencies in sequential data.

This variation in model selection helps paint a comprehensive picture of which prediction method would be superior, helping to optimize building performance and contribute to sustainable energy management practices in the built environment.

4 Results and discussions

The study adopted MSE and pseudo-R-square as the metrics chosen for our models' to further evaluate the performance of the ML models selected, ensuring robustness and providing confidince intervals around the reported performance metric. MSE is particularly advantageous for gauging the performance of non-linear models, offering insights into the accuracy and goodness of fit across a diverse set of algorithms, including XGBoost, LSTM, and SVR. Its utility lies in its ability to provide a comprehensive measure by capturing the squared differences between predicted and actual values. Pseudo R-squared is a statistical measure that evaluates the goodness of fit of a model. Unlike traditional R-squared, which applies only to linear regression models, pseudo R-squared can be used for non-linear models. For example, the MSE for XGBoost in the training set was 0.79, with a 95% confidence interval ranging from 0.75 to 0.83. Similarly, the pseudo R-squared for the XGBoost model in the testing set was 0.91, with a confidence interval of 0.88 to 0.94. These confidence intervals help assess the variability in the model's performance, highlighting the uncertainty in predictions due to the inherent variability in the data and model complexity. By incorporating these intervals, we provide a more comprehensive picture of model performance and ensure that the results are not due to random chance.

Furthermore, to ensure a fair comparison with our previous study's MLR and RF models, all models were tested using data that underwent one-hot encoding. Furthermore, all models were tuned using a grid search with the varying parameters of each model. Moreover, we employed cross-validation techniques during model training and evaluation to mitigate potential overfitting and ensure generalizability.

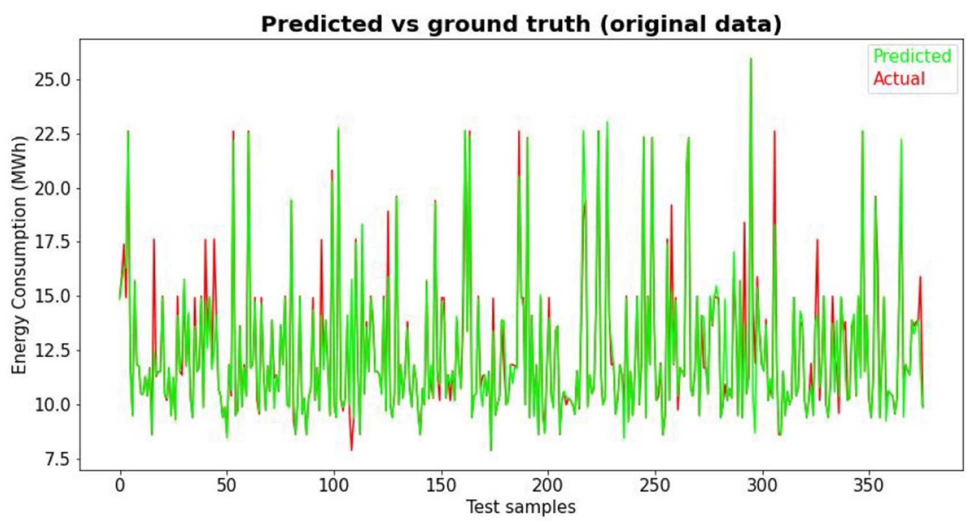
Once trained on diverse datasets, XGBoost, SVR, and LSTM models can effectively predict building envelope energy consumption across different geographical locations and climates. XGBoost excels in capturing non-linear relationships, making it adaptable to varying environmental factors like temperature and solar radiation. With its kernel flexibility, SVR handles different climates well, particularly in smaller datasets, though it remains scalable with appropriate preprocessing. LSTM is ideal for capturing temporal patterns, making it particularly suited for regions with seasonal variability.

These models are also adaptable to different building types once trained on data reflecting diverse structural, material, and usage characteristics.

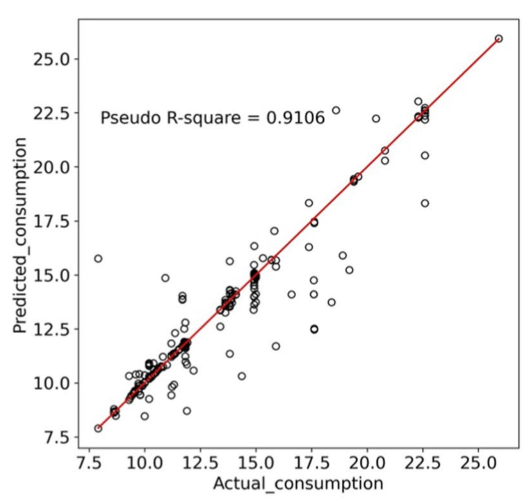
Like the RF model, the XGBoost model is adept at processing various data types, including categorical data, without requiring transformation. It was trained using 70% of the dataset and tested on the remaining 30% of the unseen data. The prediction results are displayed in Fig. 7a, where there is significant alignment between the actual and predicted values, highlighting the exceptional predictive accuracy of the XGboost model. Statistical evaluation revealed a Mean Squared Error (MSE) of 0.79 in the training set. The measure of goodness of fit, depicted in Fig. 7b, illustrates how well the explanatory variables in the XGBoost regression model explain variations in energy consumption, achieving a pseudo R-squared of 0.91 in the testing set.

Figure 8 shows the importance of the variable determined by the XGBoost model using the original data set. Feature importance quantifies the contribution of each input feature to the model's predictive accuracy, providing information on the characteristics that are most critical to predictions. It should be noted that the Y-axis represents the names of the features, while their respective importance scores are represented on the X-axis. Several features, such as thermal mass, DF summer day, U-value, and R-value, emerge as substantial contributors to the regression model.

XGBoost, a decision tree model similar to RF, exhibited an MSE value of 0.6 when tested on transformed data using one hot encoding. This performance is on par with MLR, which had an MSE of 0.6 and inferior to RF, which showed a lower



(a)



(b)

Fig. 7 Performance of XGBoost on the testing set with the original data. **a** predicted versus actual energy consumption. **b** R-square goodness of fit test

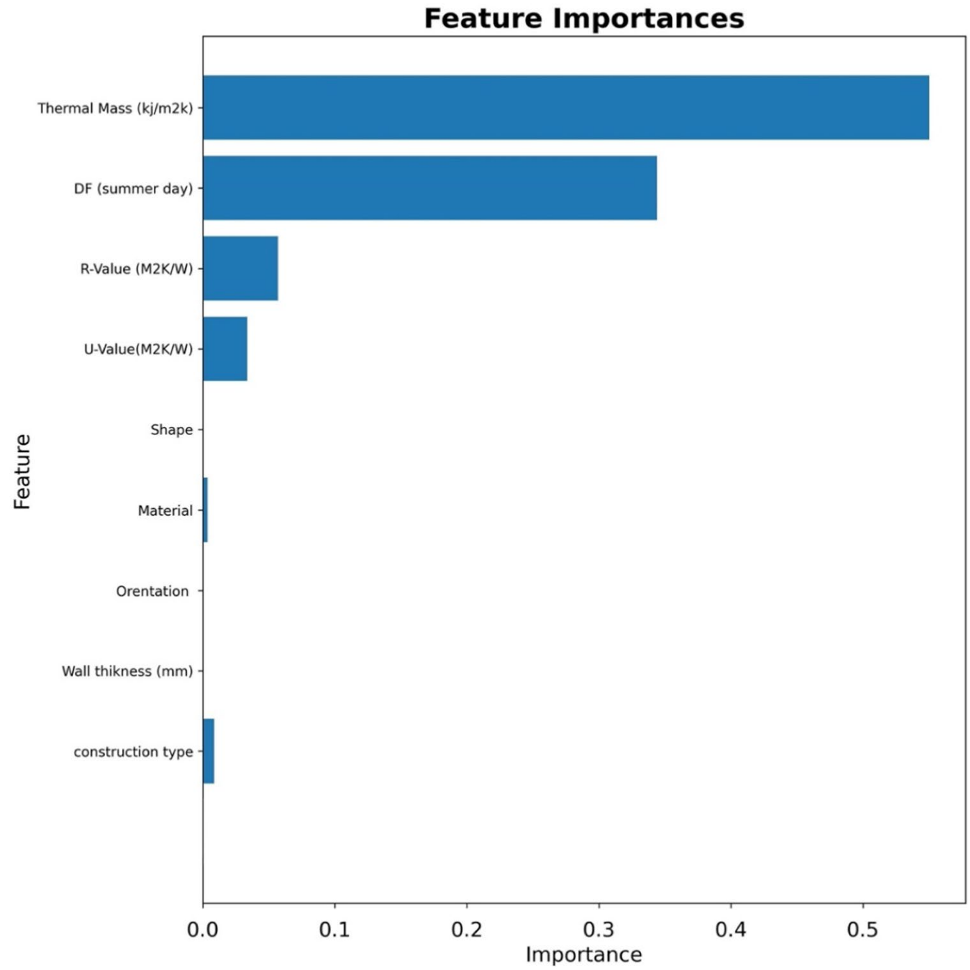


Fig. 8 Feature importance of the XGBoost model for the original data representation (without one-hot encoding)

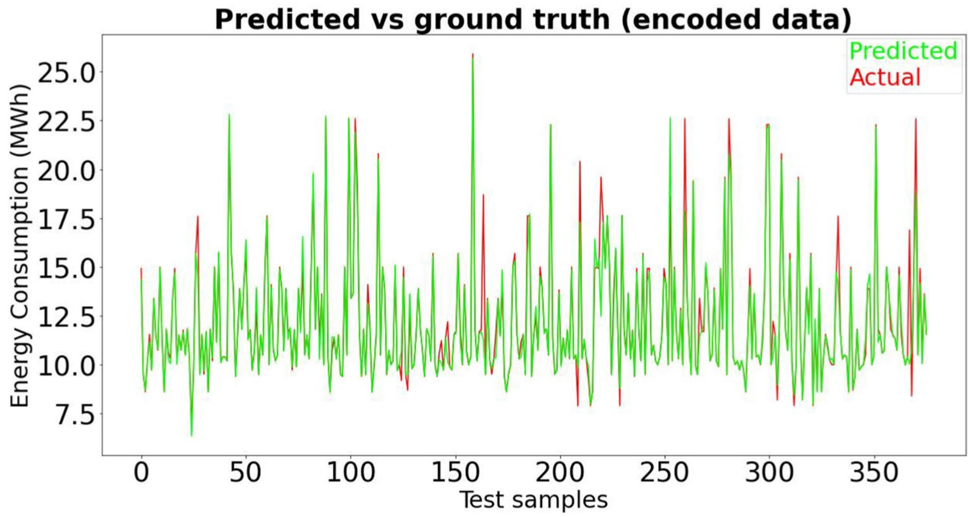
MSE of 0.06. Figure 9a illustrates the XGBoost model's predictive performance in the dataset's unseen testing portion. Notably, there are recognizable variations between ground truth (red line) and predicted values (green line), indicating challenges in the prediction accuracy of the XGBoost model on unseen testing data. The difficulty of XGBoost in accurately forecasting energy consumption values may be due to its inherent complexity. XGBoost tends to have more parameters to tune, and this complexity could lead to overfitting, especially when the dataset is small. With its simpler structure and built-in randomness, RF may be more suitable for smaller datasets and less prone to overfitting. Figure 9b shows a pseudo R-squared value of 0.94, emphasizing the ability of XGBoost to explain the variability in energy consumption, demonstrating its performance compared to traditional linear regression techniques such as MLR.

Figure 10 represents the importance of variables derived from the XGBoost model for encoded data. Likewise, in Figs. 10 (a and b), the XGBoost model, when applied to the original data representation, highlighted the significance of several variables in the regression model, remarkably thermal mass, DF summer day, U-value, and R-value as key contributors.

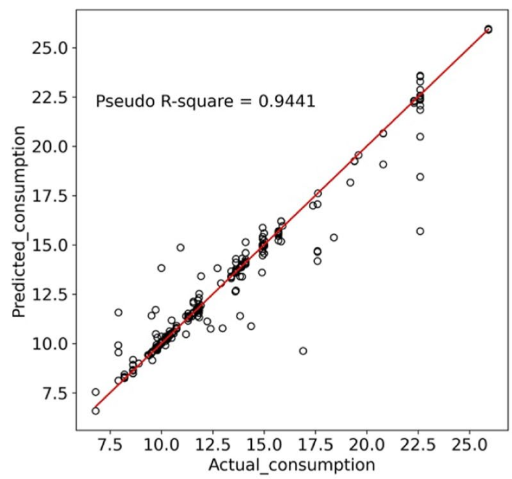
In contrast, the SVR model slightly outperformed XGBoost, achieving an MSE of 0.5. This highlights the varying predictive capabilities of different algorithms on the given data. With its distinct approach, SVR showcased its efficacy in this specific context.

Figure 11a depicts the predictive performance of the SVR on the unseen portion of the dataset. In particular, significant variations are observed between the ground truth (red line) and predicted values (green line), indicating suboptimal prediction accuracy of the SVR in unseen data.

Further analysis shows that these poor results could have been caused by outliers in the dataset, which can significantly affect the performance of the SVR model in forecasting energy consumption values. The obtained pseudo R-squared value for SVR was 0.462 in Fig. 11b, marking the lowest among the models evaluated. This outcome



(a)



(b)

Fig. 9 XGBoost performance on the test set with transformed data. **a** predicted versus actual energy consumption. **b** R-square goodness of fit test

suggests a comparatively weaker explanatory power of SVR to capture the variability in energy consumption compared to the other models examined in this study.

Figure 12 illustrates the importance of the permutation obtained from the SVR model for the encoded data. Permutation importance assesses the impact of shuffling values of each input feature on the model's predictive performance. Features that lead to a significant drop in performance when permuted are considered essential, indicating their influence on the model's predictions. The y-axis indicates the'features' names, while their corresponding importance scores are displayed on the x-axis. The analysis reveals several notable features, including thermal mass, DF summer day, U value, and R-value, which contribute to the predictive performance of the regression model. This corresponds to the XGBoost findings on the importance of features, as the four characteristics identified as the most significant were consistent.

LSTM marks a significant shift from traditional decision tree models, being part of RNNs commonly used in deep learning. Figure 13a depicts the prediction performance of the LSTM model on the unseen dataset segment, with a low MSE of 0.004, outperforming all other models utilized in this study. The minimal deviations between the predicted values (green line) and the ground truth (red line) highlight the robust predictive accuracy of LSTM in unseen data. LSTM networks excel in capturing long-term dependencies in sequential data, allowing them to retain information over extended time intervals. This unique attribute makes them suitable for forecasting tasks where past observations significantly influence future predictions. In Fig. 13b, the goodness-of-fit measure for the LSTM model is depicted, revealing its ability to explicate the variance in energy consumption. The LSTM model achieved a remarkable pseudo-R-squared value of 0.993

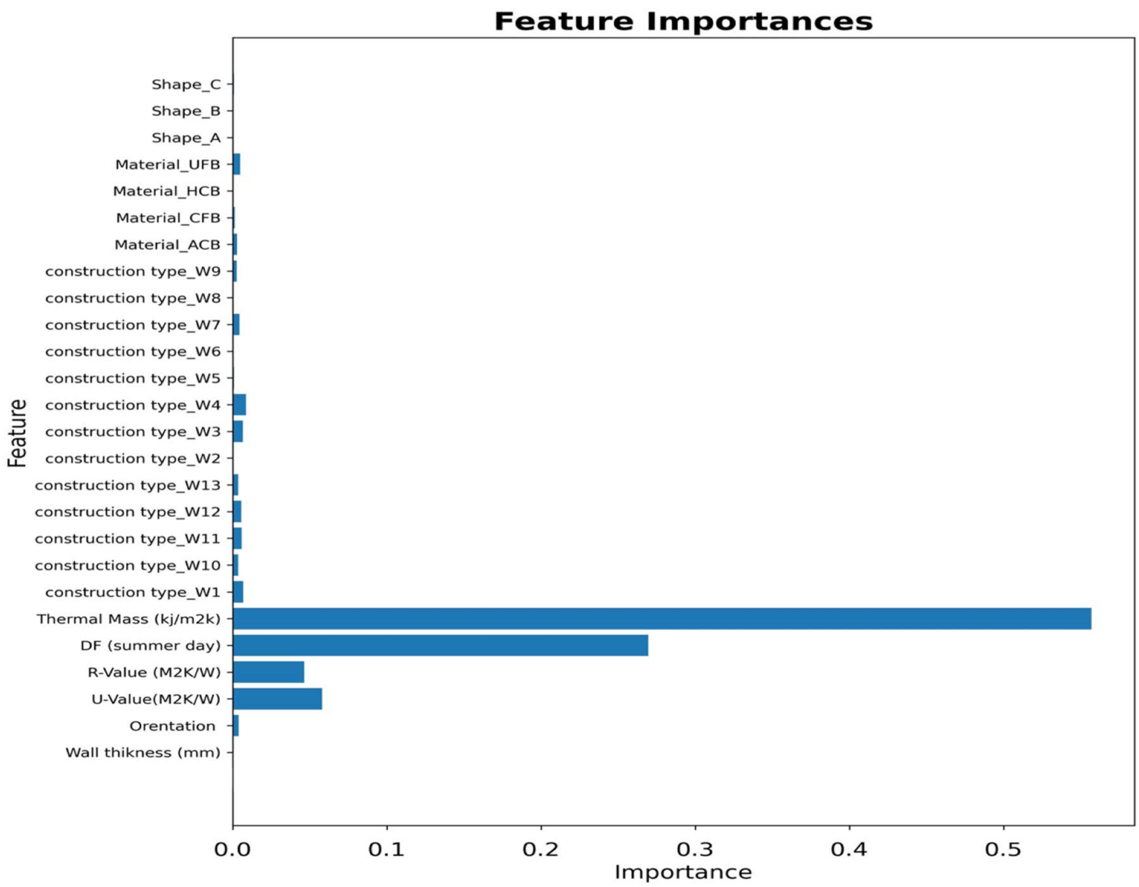


Fig. 10 Feature importance of the XGBoost model for the transformed data representation (including one-hot encoding)

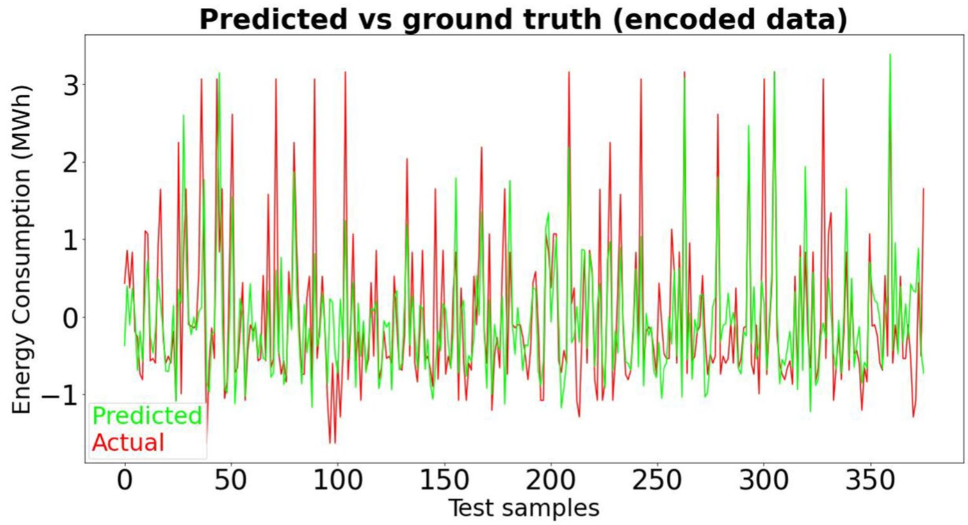
in the testing set. The high pseudo-R-squared value indicates 'LSTM's ability to capture and explain variations in energy consumption, outperforming other models in this study.

These findings reaffirm the superiority and reliability of''LSTM' in energy consumption prediction tasks, establishing it as a key model for optimizing the building envelope design. Figure 14 represents the perturbation analysis results obtained from the LSTM model for the encoded data. Perturbation analysis involves systematically perturbing the values of input features and observing the effect on the model's predictions. Unlike permutation importance, which evaluates the impact of shuffle feature values, perturbation analysis directly modifies the feature values while keeping the rest of the data unchanged. On the y-axis, the'features' names are listed, while the x-axis represents the change in the model performance metric resulting from perturbing each feature. This analysis highlights significant features such as wall thickness, which play crucial roles in shaping the predictive performance of the regression model. Furthermore, discerning the attribute significance for discrete level factors, such as construction shapes, is straightforward. Conspicuously, construction shape A holds greater significance in model prediction.

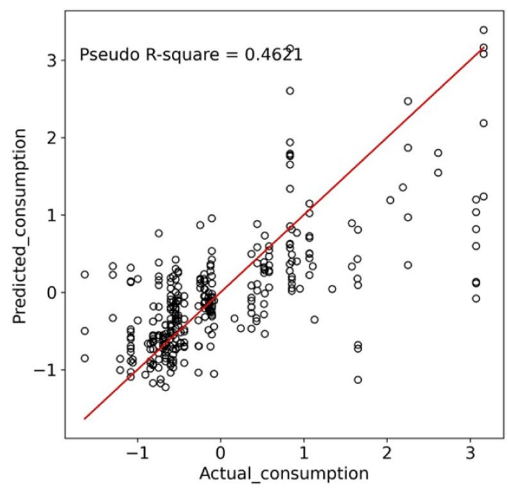
As indicated in Table 1, various ML models have been employed for building energy consumption prediction. The researchers in [16] evaluated energy consumption and building envelope performance, identifying GB as suitable with an R-squared of 0.994. In [23], heat fluxes through a novel active building envelope (TABE) were predicted, with XGBoost achieving an R-squared of 0.82. In [14], the researchers predicted the energy consumption of PCM-integrated residential buildings, achieving high reliability with SVM and ANN models. Our proposed model achieved an R-squared of 0.993, demonstrating comparable predictive accuracy. The differences in performance could be attributed to variations in the dataset, such as the building types, environmental factors, and data preprocessing techniques used in each study. The choice of algorithm also plays a crucial role in performance, as models like XGBoost and LSTM excel in capturing non-linear relationships and temporal patterns, respectively.

Finally, While the LSTM model demonstrated impressive performance in energy consumption predection with a high pseudo R-squared value of 0.993, it is important to acknowledge its limitations. LSTM excels in capturing temporal

Fig. 11 SVR performance on the test set with transformed data. **a** predicted versus actual energy consumption. **b** R-square goodness of fit test



(a)



(b)

dependencies and long-term patterns, making it highly suitable for datasets with significant sequential or time-series components. However, its performance may not be as robust in scenarios with limited temporal data or where other models, such as XGBoost or RF, can better handle non-linear relationships in the absence of strong temporal patterns. Additionally, LSTM models require careful tuning and are computationally more expensive than simpler models, which may limit their practical application in certain contexts. Therefore, while LSTM offers significant advantages in specific use cases, it should be carefully evaluated against other models based on the nature and complexity of the data at hand.

5 Conclusions

This research aimed to build on previous research in which a thermal performance prediction model was developed for wall envelopes in the UAE. To this end, the previous research used ML methods, specifically MLR and RF, which show the superiority of RF over MLR. This research expands on these findings by testing other ML models, XGBoost, SVR, and LSTM, comparing their performance to RF, MLR, and other models from the literature.

After extensive simulations to evaluate the predictive capabilities of each model, LSTM emerged as the superior predictor, even outperforming RF. This demonstrates the potential of LSTM and highlights its superiority over well-established methods such as RF. The ability LSTM to achieve higher accuracy in predicting thermal performance

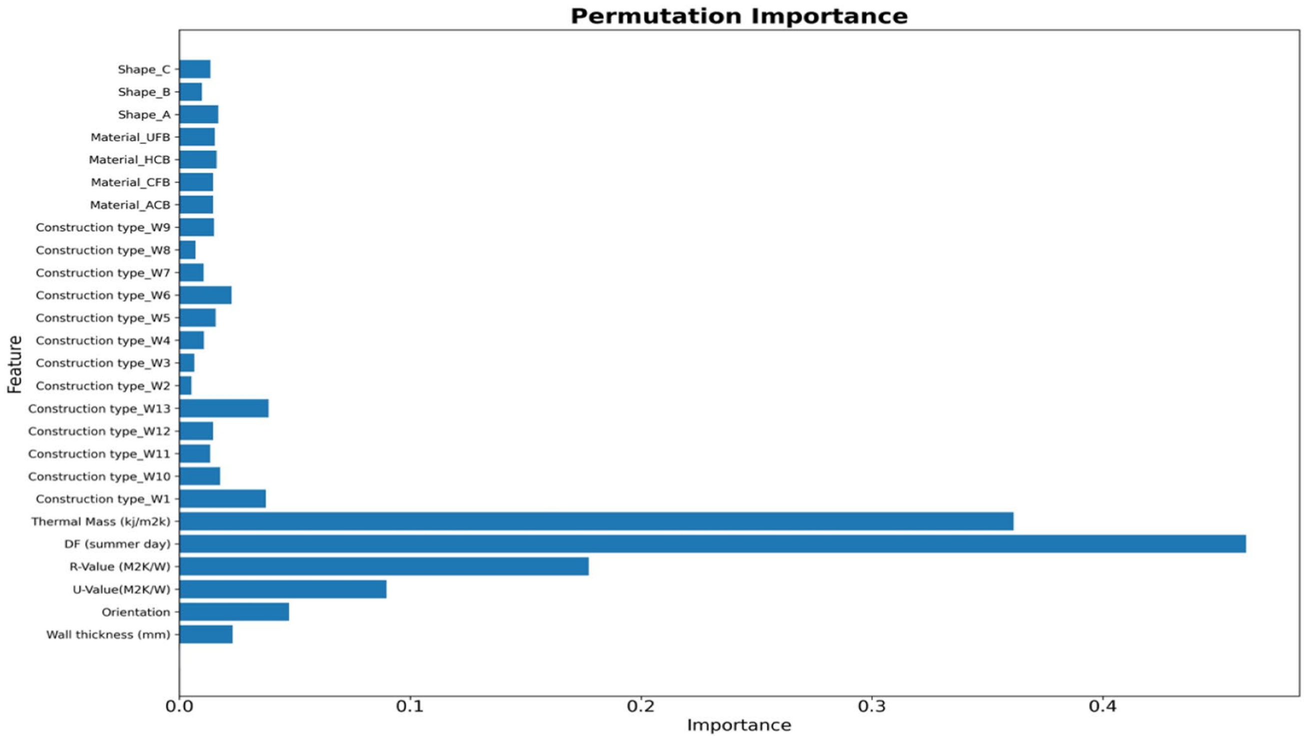
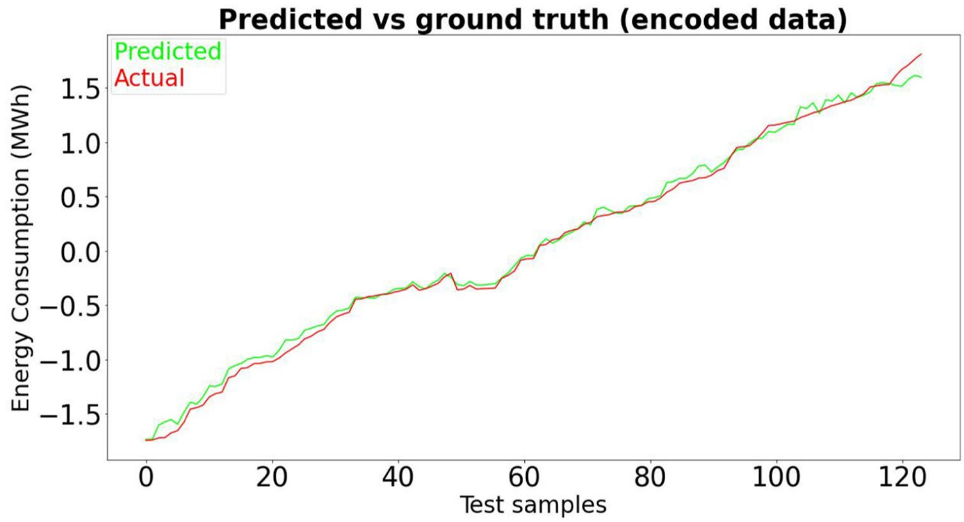


Fig. 12 Feature importance of the SVR model for the transformed data representation (including one-hot encoding)

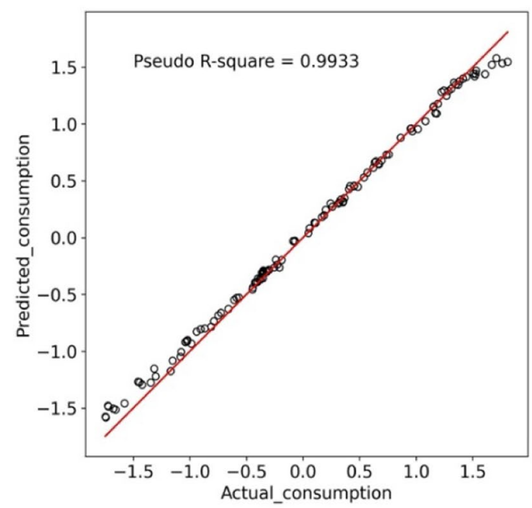
further solidifies its position as a cutting-edge tool for enhancing energy efficiency in building construction. This research broadens the understanding of ML methodologies in energy performance prediction and paves the way for the use of techniques such as LSTM to address critical challenges in the design and operation of sustainable buildings.

Our study findings contribute valuable insights and input especially in hot climates as follows:

- Our current study builds on the foundation of previous research published by the author, proposing a robust ML framework as a powerful tool for the construction industry to predict energy performance and raise the energy efficiency standards of buildings, particularly in challenging environments such as hot climates where energy optimization is critica
- Previously, RF was acknowledged as an effective method to explain energy consumption patterns in the building sector. However, our latest findings reveal LSTM as a transformative force in predictive modeling. The exceptional performance of LSTM highlights its potential to revolutionize energy performance prediction methodologies, offering unparalleled accuracy and insight. This is especially beneficial in regions with extreme temperature variations where accurate forecasting of energy needs is essential.
- Rigorous statistical tests were conducted on the prediction results, revealing significant variation in model performance. Although RF exhibited strong predictive power with an R-squared value of 0.95 in the previous study, MLR showed comparatively weaker performance with an R-squared value of 0.42. In our expanded investigation, SVR and XGBoost also fell short, particularly in comparison to the unequaled performance of LSTM. SVR achieved an R-squared value of 0.462, and XGBoost an R-squared value of 0.94. Furthermore, XGBoost had an MSE value of 0.6, while SVR achieved an MSE of 0.5. On the contrary, LSTM emerged as the main performer, delivering unprecedented precision with an R-squared value of 0.993 and an MSE of 0.004, reaffirming its position as the optimal choice for the prediction of energy performance in building envelopes.



(a)



(b)

Fig. 13 LSTM performance in the testing set with transformed data. **a** predicted versus actual energy consumption. **b** R-square goodness of fit test

- This work bridges a critical gap in the literature by linking theoretical ML models with practical applications in building performance simulations. Provides actionable insights that can be applied directly in real-world scenarios, enhancing the practical value of academic research in building technology.
- This study significantly advances sustainable building practices by providing architects and engineers with a reliable tool. Accurate energy consumption prediction empowers professionals to make informed decisions regarding building designs, materials, and technologies, thereby driving the construction industry towards higher energy efficiency standards.

The findings of both studies contribute to a growing body of research that influences ML for energy efficiency in buildings. While RF remains a robust and interpretable method, the advent of LSTM introduces unparalleled precision, setting new standards for predictive modeling in sustainable building design. Together, these works underscore the transformative potential of ML in driving energy efficiency and sustainability in the construction industry, particularly in challenging environments such as hot climates where energy optimization is crucial.

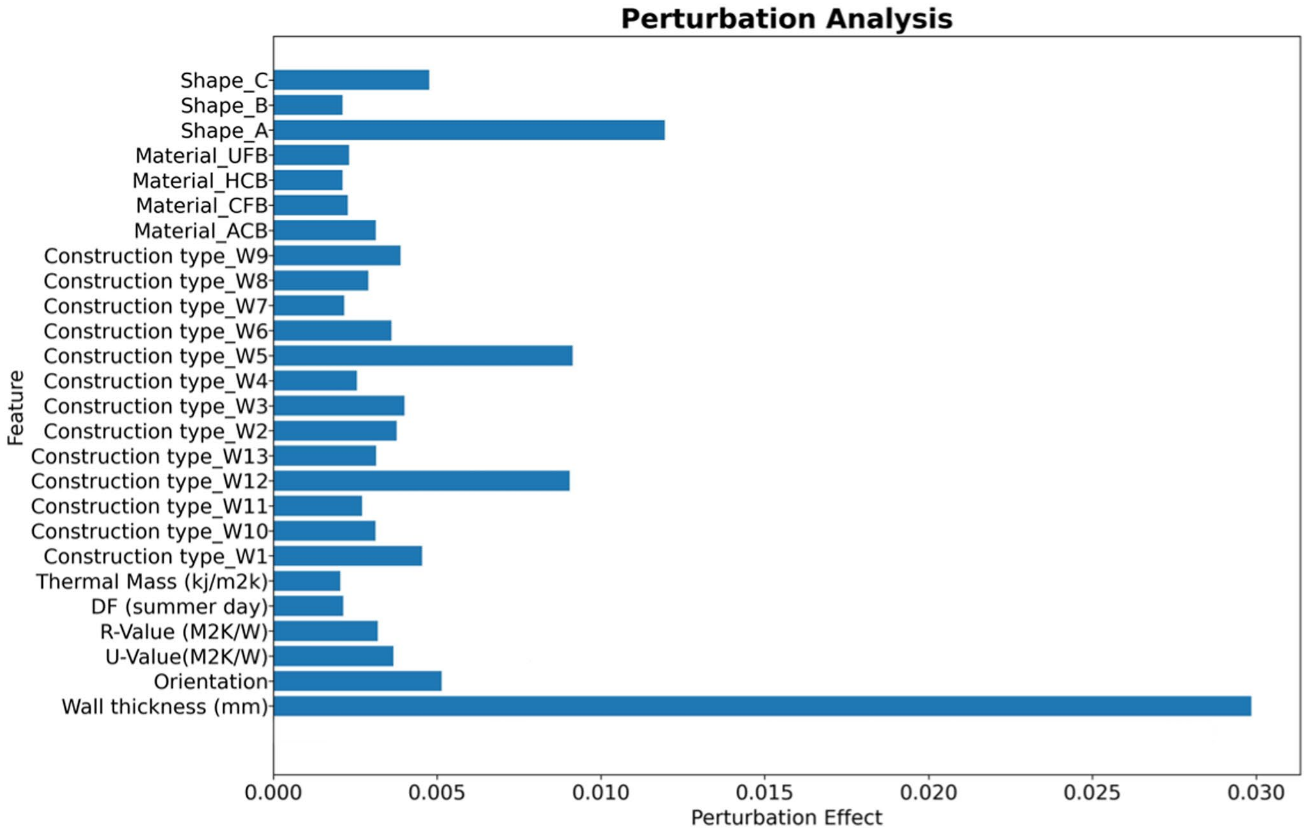


Fig. 14 Attribute importance of the LSTM model for the transformed data representation (including one-hot encoding)

Author contributions Conceptualization: AH, AM Methodology, AH, and AD Investigation, TB, AH Supervision, TB, AH Writing original draft, AH, AM, AA Writing review & editing: AH, TB.

Data availability The datasets generated during the current study are available from the corresponding author upon reasonable request.

Declarations

Competing interests The authors declare no competing interests.

References

1. López-Guerrero RE, Cruz AS, Hong T, Carpio M. Optimizing urban housing design: improving thermo-energy performance and mitigating heat emissions from buildings–a Latin American case study. Urb Clim. 2024;57: 102119.
2. Fei W, et al. The critical role of the construction industry in achieving the sustainable development goals (SDGs): delivering projects for the common good. Sustainability. 2021. https://doi.org/10.3390/su13169112.
3. Huang L, et al. Carbon emission of global construction sector. Renew Sustain Energ Rev. 2018;81:1906–16.
4. Jannat N, Hussien A, Abdullah B, Cotgrave A. A comparative simulation study of the thermal performances of the building envelope wall materials in the tropics. Sustainability. 2020;12(12):4892.

5. Cruz AS, Caldas LR, Mendes VM, Mendes JC, Bastos LEG. Multi-objective optimization based on surrogate models for sustainable building design: a systematic literature review. Build Environ. 2024. https://doi.org/10.1016/j.buildenv.2024.112147.
6. Hafez F, et al. Energy efficiency in sustainable buildings: a systematic review with taxonomy, challenges, motivations, methodological aspects, recommendations, and pathways for future research. Energ Strat Rev. 2023;45: 101013.
7. Jia L, et al. Interaction between thermal comfort, indoor air quality and ventilation energy consumption of educational buildings: a comprehensive review. Buildings. 2021;11(12):591.
8. Ullah Z, Al-Turjman F, Mostarda L, Gagliardi R. Applications of artificial intelligence and machine learning in smart cities. Comput Commun. 2020;145:313–23.
9. Mukhamediev R, et al. Review of artificial intelligence and machine learning technologies: classification, restrictions, opportunities, and challenges. Mathematics. 2022;10(15):2552.
10. Dunjko V, Briegel H. Machine learning and artificial intelligence in the quantum domain. arXiv Prepr. 2017. https://doi.org/10.1088/1361-6633/aab406.
11. Zhang L, et al. A review of machine learning in building load prediction. Appl Energ. 2021;285: 116452.
12. Hussien A, et al. Predicting energy performances of buildings' envelope wall materials via the random forest algorithm. J Build Engin. 2023;69: 106263.
13. Alvarez-Sanz M, et al. Ranking building design and operation parameters for residential heating demand forecasting with machine learning. J Build Engin. 2024;86: 108817.
14. Zhussupbekov M, et al. Forecasting energy demand of PCM integrated residential buildings: a machine learning approach. J Build Engin. 2023;70: 106335.
15. Seyedzadeh S, Rahimian F, Rastogi P, Glesk I. Tuning machine learning models for prediction of building energy loads. Sustain Cities Soc. 2019;47: 101484.
16. Long L, et al. Machine learning-based predictive models for energy consumption estimation in energy-efficient building envelope design. Int Conf Sustain Civil Engin Archit. 2023. https://doi.org/10.1007/978-981-99-7434-4_58.
17. Tian Z, Shi X, Hong S. Exploring data-driven building energy-efficient design of envelopes based on their quantified impacts. J Build Engin. 2021;42: 103018.
18. Magoulès F, Piliougine M, Elizondo D. Support vector regression for electricity consumption prediction in a building in japan. In: Magoulès F, Piliougine M, Elizondo D, editors. 2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES). New York: IEEE; 2016. p. 189–96.
19. Zhong H, et al. Vector field-based support vector regression for building energy consumption prediction. Appl Energ. 2019;242:403–14.
20. Chen J, Lin J. Developing an SVM-based risk hedging prediction model for construction material suppliers. Autom Constr. 2010;19(6):702–8.
21. Jaber F, Al-Zwainy F, Hachem S. Optimizing predictive performance for construction projects utilizing support vector machine technique. Cogent Engin. 2019;6(1):1685860.
22. Peško I, et al. Estimation of costs and durations of construction of urban roads using ANN and SVM. Complexity. 2017;2017:2450370.
23. Shen Z, et al. Machine learning–assisted prediction of heat fluxes through thermally anisotropic building envelopes. Build Environ. 2023;234: 110157.
24. Kim C, Kim M, Song Y. Sequence-to-sequence deep learning model for building energy consumption prediction with dynamic simulation modeling. J Build Engin. 2021;43: 102577.
25. Kolay E, Baser T. Estimating the dry unit weight of compacted soils using general linear model and multi-layer perceptron neural networks. Appl Soft Comput. 2014;18:223–31.
26. Barbaresi A, et al. Application of machine learning models for fast and accurate predictions of building energy need. Energies. 2022;15(4):1266.
27. Archibong-Eso EU, Enyia JD. Machine learning models for energy prediction in a low carbon building. In: Archibong-Eso EU, Enyia JD, editors. SPE Nigeria annual international conference and exhibition. Lagos: Society of Petroleum Engineers; 2023.
28. Liu T, et al. Study on deep reinforcement learning techniques for building energy consumption forecasting. Energ Build. 2020;208: 109675.
29. Wen L, Zhou K, Yang S. Load demand forecasting of residential buildings using a deep learning model. Electr Power Syst Res. 2020;179: 106073.
30. Kahloot K, Ekler P. Algorithmic splitting: a method for dataset preparation. IEEE Access. 2021;9:125229–37.
31. Wade C, Glynn K. Hands-on Gradient Boosting with XGBoost and Scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python. 1st ed. Birmingham: Packt Publishing; 2020.
32. Hearst M, Dumais S, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Their Appl. 1998;13(4):18–28.