Artificial Intelligence Review (2025) 58:211
https://doi.org/10.1007/s10462-025-11226-6

Fast machine learning for building management systems

Mohammed Mshragi[1] · **Ioan Petri**[1]

Accepted: 4 April 2025

Abstract
Building management systems (BMSs) are increasingly integrating advanced machine learning (ML) and artificial intelligence (AI) capabilities to enhance operational efficiency and responsiveness. The transformation of BMSs involves a wide range of environmental, behavioural, economical and technical factors as well as optimum performance considerations in order to reach energy efficiency and for long term sustainability. Existing BMSs can only provide local adaptability by creating and managing information for a built asset lacking the capability to learn and adapt based on performance objectives. This research provides a comprehensive review of ML techniques in BMSs, with particular emphasis and demonstration of fast machine learning (FastML) techniques in a real-case study application. The study reviews optimization methods for ML algorithms, focusing on Long Short-Term Memory (LSTM) networks for energy consumption forecasting and exploring solutions that leverage hardware accelerators for low-latency and high-throughput processing. The High-Level Synthesis for Machine Learning (HLS4ML) framework facilitates deployment of fast machine learning models with BMSs, achieving substantial gains in hardware efficiency and inference speed in resource-constrained environments. Findings reveal that HLS4ML-optimized models maintain accuracy while offering computational efficiency through techniques like pruning and quantization, supporting real-time BMS applications. This research significantly contributes to the development of intelligent BMSs by integrating ML algorithms with advanced hardware solutions, ultimately improving energy management, occupant comfort, and safety in modern buildings.

Keywords Fast machine learning · Building management systems · Energy forecasting · High level specification languages · Building automation

Abbreviations

ANN	Artificial neural network
AI	Artificial intelligence

✉ Mohammed Mshragi
mshragim@cardiff.ac.uk

[1] School of Engineering, Cardiff University, Cardiff, UK

Published online: 17 April 2025

BMS	Building management system
BEMS	Building energy management system
BAS	Building automation system
BIM	Building information modeling
CNN	Convolutional neural network
FL	Federated learning
HFL	Horizontal federated learning
VFL	Vertical federated learning
FDD	Fault detection and diagnosis
FPGA	Field-programmable gate array
HLS4ML	High-level synthesis for machine learning
HVAC	Heating, ventilation, and air conditioning
IoT	Internet of Things
LLM	Large language model
LSTM	Long short-term memory
ML	Machine learning
MPC	Model predictive control
PCA	Principal component analysis
RL	Reinforcement learning
RNN	Recurrent neural network
SSL	Semi-supervised learning
TL	Transfer learning
NILM	Non-intrusive load monitoring
LDA	Linear discriminant analysis
t-SNE	t-Stochastic neighborhood embedding
SOM	Self-organizing maps
GAs	Genetic algorithms
ICA	Independent component analysis
GMM	Gaussian mixture model
IAQ	Indoor air quality
MLP	Multilayer perceptron
NAS	Neural architecture search
LR	Logistic regression
SVR	Support vector regression
AR	Auto-regressive
RF	Random Forest
XGBoost	Extreme gradient boosting
AdaBoost	Adaptive boosting
ARMA	Auto-regressive moving average
RT	Real-time
DQN	Deep Q-network
DDQN	Double deep Q-Network
SARSA	State-Action-Reward-State-Action
AR	Auto-regressive
GA	Genetic algorithm
TLD	Transfer learning domain

RCM	Resource consumption model
TL-CNN	Transfer learning convolutional neural network
TL-LSTM	Transfer learning long short-term memory
LNCS	Springer Lecture Notes in Computer Science
Q-Learning	Quality learning algorithm
DT-MPC	Decision tree model predictive control
SBNMF	Semi-binary nonnegative matrix factorization
SOM	Self-organizing maps

1 Introduction

ML for BMSs represents a transformative advancement in facility management, simplifying the operation, maintenance, and optimization of building systems. Building Automation Systems (BASs) regulate critical infrastructure aspects such as heating, ventilation, and air conditioning (HVAC), lighting, and energy consumption in buildings (Abdullah et al. 2022). Recent advancements in sensing and Internet of Things (IoT) technologies have facilitated data-driven approaches in BMSs, significantly enhancing efficiency, cost-effectiveness, and occupant comfort (Finck et al. 2018). To enhance responsiveness in dynamic built environments, traditional ML techniques in BMSs must be complemented with software and hardware accelerators. Historically, BMSs relied on rule-based control methods, limiting their ability to adapt effectively to dynamic factors such as fluctuating energy tariffs and changing meteorological conditions (Finck et al. 2018). The integration of sophisticated sensing technologies and IoT devices has ushered in a transformative era of data-driven building management, greatly improving both efficiency and occupant comfort (Abuimara et al. 2021). In contemporary settings, data-driven methodologies-especially those employing ML and artificial intelligence (AI)-are increasingly integrated into BMSs to bolster their functionality, efficiency, and responsiveness. AI-powered BMSs leverage advanced analytics, predictive modeling, and intelligent automation to optimize operations. By harnessing the extensive data generated by BMSs, these systems can uncover patterns, trends, and anomalies that traditional rule-based systems may overlook. This adaptive capacity allows BMSs to dynamically respond to fluctuating environmental conditions, optimize energy consumption, and enhance occupant safety. However, challenges persist in incorporating AI and ML into BMSs, including ensuring data quality, achieving rapid decision-making, and navigating implementation complexities (Puiu and Fortis 2024). Real-time decision-making is critical, as traditional ML algorithms often exhibit slow response times (Duarte et al. 2022a). FastML, which refers to rapid machine learning techniques that enhance performance, is becoming increasingly important in this context. Advancements in hardware, particularly in Field-Programmable Gate Arrays (FPGAs), are essential for addressing the need for rapid ML decision-making. To address these challenges, the primary objectives of this research are to (i) provide a comprehensive review of existing ML techniques within BMSs, focusing on a diverse range of algorithms and applications, and (ii) investigate the emergence and effectiveness of FastML techniques with hardware accelerators for energy management applications within BMSs, as demonstrated through a case study. This review aims to enable BMSs to swiftly adapt to changing environmental conditions, ensuring occupant safety and comfort while optimizing energy usage. In the subsequent sections, we will

explore existing studies on ML-based BMSs, exploring the methodologies employed and specific techniques for optimizing LSTM models for energy management. The case study will highlight a comparative analysis of inference speeds across various LSTM models, providing insights into the practical applications of these techniques. This exploration ultimately seeks to deliver a comprehensive understanding of current advancements and future directions in this rapidly evolving field, highlighting the implications for research and practice in BMS solutions.

1.1 Existing studies on ML-based BMSs

Numerous studies have explored the application of AI and ML in BMSs, with a focus on various aspects such as power consumption, anomaly detection, occupants' satisfaction, and security. Table 1 provides an overview of the existing research in this field. One notable study by Mazhar et al. (2022) delves into the integration of 5G technology into smart building management systems (BMSs). The research emphasizes the need for sustainable solutions in the face of resource constraints and population growth. It advocates for the incorporation of intelligent systems within smart homes, leveraging IoT and cloud technologies to address challenges across different domains. The study highlights the importance of IoT-enabled energy-conserving buildings and calls for increased awareness and financial incentives, particularly in commercial settings. Additionally, it explores how 5 G can enhance service quality, network capacity, and AI integration in automated systems while addressing privacy concerns. The research provides valuable insights into advancing smart city evolution within the context of big data and 5 G advancements, considering challenges like building

Table 1 Overview of studies on ML applications in BMSs

Authors	Year	Focus area	Main findings	Key contributions	Challenges addressed
Mazhar et al.	2022	Integration of 5 G in smart building management	Need for sustainable solutions, 5 G's potential in automation and privacy concerns	Incorporation of intelligent systems within smart homes, leveraging IoT and cloud technologies	Building penetration issues
Himeur et al.	2023	AI-big data analytics in BAMSs	Importance of machine learning, challenges in security and scalability	Supervised, unsupervised, semi-supervised, reinforcement learning	Security, interoperability, scalability
Digitemie and Ekemezie	2024	Building Energy Management Systems (BEMS)	BEMS' role in energy efficiency, challenges, and future prospects	Utilization of sensors, controllers, and networks	Costs, integration issues
Heidari et al.	2024	Integration of BIM and AI in construction	Potential revolution in construction, challenges in integration	Leveraging machine learning algorithms and smart devices	Data integration, software compatibility
Ngo et al.	2024	Cloud-based AI system for energy management	Effectiveness of the system in energy monitoring and prediction	Combination of cloud technology and AI algorithms	Long-term energy prediction
Chen et al.	2023	Interpretable ML techniques in building energy management	Challenges, future opportunities in interpretable ML	Ante-hoc, post-hoc approaches	Terminology confusion, limited techniques

This table summarizes key contributions, including authors, publication year, focus areas, main findings, and challenges

penetration issues and existing structures. Another comprehensive study by Himeur et al. (2023) focuses on the application of AI-big data analytics in Building Automation and Management Systems (BAMSs). The study examines various functionalities, including energy prediction, fault detection, anomaly spotting, and indoor environment evaluation. Different AI models, such as supervised, unsupervised, semi-supervised, and reinforcement learning, were tested. The research highlights the significance of machine learning, IoT, and connectivity in shaping BAMSs. It acknowledges the effectiveness of supervised learning with labeled data and the promise shown by unsupervised learning despite lower efficiency. The study also emphasizes the need to address challenges such as security, interoperability, and scalability. Furthermore, Building Energy Management Systems (BEMS) play a vital role in improving energy efficiency and sustainability in buildings. They oversee and regulate systems like HVAC and lighting using components such as sensors, controllers, and networks to collect data and optimize energy usage. Despite challenges like costs and integration, BEMS offer advantages such as lower energy consumption and enhanced comfort for occupants (Digitemie and Ekemezie 2024). Technological advancements like IoT and AI are addressing these challenges, making BEMS more accessible and efficient. The integration of AI and machine learning holds promise for further improving energy-saving capabilities and building performance. In summary, BEMS are crucial for achieving energy efficiency and sustainability goals, delivering significant savings and environmental benefits (Digitemie and Ekemezie 2024). As illustrated in Fig. 1, various ML techniques and model optimizations are employed to achieve these enhancements. Nevertheless, despite the substantial advantages of incorporating AI and ML into BMSs, several challenges persist.

A systematic review conducted by Heidari et al. (2024) explores the integration of Building Information Modeling (BIM) and Artificial Intelligence (AI) in the construction industry. This integration has the potential to revolutionize the sector by enhancing decision-making, optimizing processes, and increasing overall efficiency. By leveraging machine learning algorithms and smart devices, AI can enhance BIM's capabilities, includ-

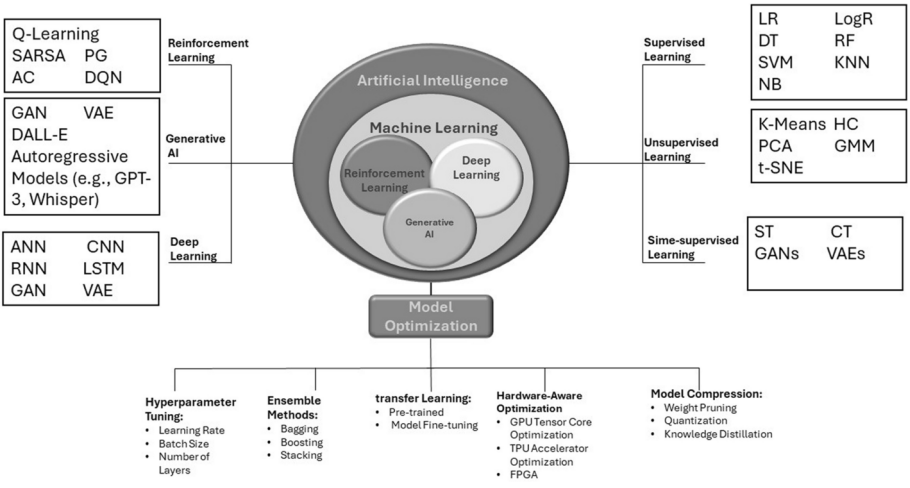

Fig. 1 Overview of ML techniques and optimization strategies. This diagram categorizes AI methods-including supervised, unsupervised, and semi-supervised learning, deep learning, generative AI, and reinforcement learning alongside optimization strategies like model compression, highlighting their applications in BMS

ing predicting building performance, identifying design issues, and optimizing construction processes. However, successful integration requires careful consideration of factors such as data integration and software compatibility. Further research is needed to address the interoperability and scalability challenges of integrating BIM and AI. Nonetheless, the potential benefits, including improved energy efficiency and accurate cost estimation, make this integration a promising direction for the future of construction. Additionally, a survey conducted by Ngo et al. (2024) introduces a cloud-based AI system for managing energy in buildings. This system combines cloud technology, AI algorithms, optimization methods, and web applications to collect, analyze, and visualize energy consumption data. It consists of three layers: the data layer for storing energy-related data, the AI-based analytics layer for processing and predicting energy usage, and the decision-support information layer for presenting insights and interactive visualization. Practical case studies were conducted to test the system's effectiveness in monitoring and predicting energy consumption while providing useful information for building managers and users. This study contributes to the knowledge of energy efficiency in buildings and offers a valuable tool for implementing smart energy management systems. Future research can further explore long-term. Furthermore, Chen et al. (2023c) provide a comprehensive review of previous research on interpretable machine learning (ML) techniques in building energy management. The article categorizes the applications into ante-hoc and post-hoc approaches and highlights challenges such as terminology confusion and limited techniques. The article suggests future opportunities, including exploring interpretable ML for classification tasks and developing customized models for different users. Availability of open datasets and interpretable deep reinforcement learning models are also proposed. Overall, these studies highlight the importance of AI, ML, and technological advancements like IoT and 5 G in the field of building management systems. They provide insights into the potential applications, challenges, and future directions for creating more intelligent and efficient buildings.

1.2 Objectives and scope

This paper aims to (i) provide a comprehensive review of existing machine learning (ML) techniques in BMSs (BMS) and (ii) investigate the emergence of FastML techniques for BMSs applications with a case study example. The review focuses on ML methods applied to various built asset types, including residential, commercial, and industrial buildings.

The section on "AI Applications in BMS" examines how advanced AI analytics enhance BMS functionality through:

FastML and optimization: The paper evaluates the effectiveness of various ML methods in optimization tasks, such as energy management, predictive maintenance, and resource allocation. It discusses how ML models leverage abundant building data to streamline operations and improve decision-making.

ML and AI in BMS applications: This review highlights state-of-the-art ML-powered BMS solutions and identifies promising research avenues, with a focus on areas like fault detection, occupant behavior analysis, and data contextualization to improve BMS performance. Optimization techniques, including pruning and quantization, are also analyzed, demonstrating the suitability of quantized models for computationally

efficient, real-time applications. This approach aims to enhance BMS functionality, efficiency, and sustainability on both individual and city-wide scales.

Key research questions: This investigation addresses the following questions:

1. How Fast Machine Learning can improve building management systems performance, automation and efficiency with identification of gaps in research and development to be addressed to advance ML-powered BMS solutions?
2. What are the key ML methods utilised for building performance management and optimization tasks, including energy management and predictive maintenance?
3. How to create, deploy and test a high language specification fast machine learning model using an energy forecasting application from a real building case study?

This paper is organized as follows: Sect. 1.3 describes the methodology, Sect. 2 explores applications of ML in BMS and optimization in BMS using ML with research gaps and best practices. Section 3 provides the evaluation of this work and Sect. 4 reports relevant discussions around the findings. Section 5 presents the conclusions of this research. Figure 2 illustrates the integration of BMS and ML, providing a visual representation of the concepts discussed. Through this comprehensive overview, we explore how ML-driven BMS solutions impact various BMS applications and suggest directions for further advancements.

1.3 Methodology

This review examines the available research on ML applications in BMS through a three-stage methodology.

Stage 1: Planning

The first stage involved defining the review's scope, research questions, and target databases, which included Google Scholar, ACM Digital Library, IEEE Xplore Digital Library, and Springer Lecture Notes in Computer Science (LNCS). The review focused on studies published between 2016 and 2024 and was managed using EndNote reference software.

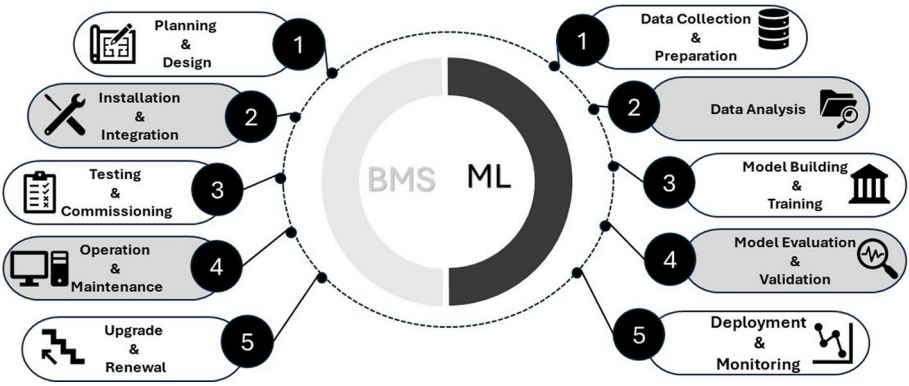

Fig. 2 Integration of BMS and ML. This diagram illustrates the key phases, including planning, installation, data collection, analysis, model building, evaluation, and deployment

Stage 2: Searching and Filtering

In the second stage, a systematic search process was established. Key terms and interchangeable terms were defined for use in the search, summarized in Table 2). Initial screening involved reviewing titles and abstracts. Inclusion criteria included English language, document type (full-text, conference/journal papers, or books), and publication year (2016–2024). After removing duplicates, the screened papers were reviewed in full. Studies that did not address ML techniques in BMS or failed to present primary research findings were excluded, resulting in 147 selected papers imported into EndNote.

Stage 3: Evaluation and Extraction

The final stage involved evaluating the quality and relevance of the selected articles. Papers were assessed based on three criteria:

1. **Clarity of methodology:** Whether study methods were clearly described and understandable.
2. **Provision of results:** How well the studies presented outcomes and supporting data.
3. **Relevance to research questions:** How closely each study aligned with the review's research questions.

After evaluation, important details were extracted, including: ML models used, Optimal models identified,Targeted BMS applications, Scale of BMS implementation (e.g., individual buildings, entire cities).

This review, as shown in Fig. 3, encompasses a wide range of ML applications within BMS, examining each in terms of motivations, constraints, and methods. The findings contribute to understanding how ML enhances BMS operations by automating tasks, improving decision-making, and optimizing energy consumption and operational costs. Through analyses of applications such as fault detection, predictive maintenance, energy forecasting, and anomaly detection, this paper provides insights into a more sustainable BMS framework. Furthermore, the inclusion of FastML techniques, such as model quantization and pruning, proves critical in enabling real-time applications essential for energy management and occupant comfort. This discussion not only focuses on the present state of ML in BMS but also lays out a trajectory for future research and development in this rapidly evolving domain. The final sections discuss existing research gaps, especially concerning FastML's scalabil-

Table 2 Search terminology for BMS and ML

Main terminology	Search terminology
Building management system	Building automation system, HVAC-based ML, building energy management system (BEMS), building information modeling (BIM)
Machine Learning Application	Predictive Maintenance, Fault Detection and Diagnosis (FDD), Energy Forecasting, Occupancy Detection and Prediction, Thermal Comfort Optimization, Load Shifting, Demand Response, Anomaly Detection, Reinforcement Learning for Controls
Fast Machine Learning	HLS4ML, QONNX
Optimization	Model Quantization, Hyperparameter Optimization, Model Compilation, Parallel Processing, Pruning Techniques

This table presents main terms alongside relevant search terminology

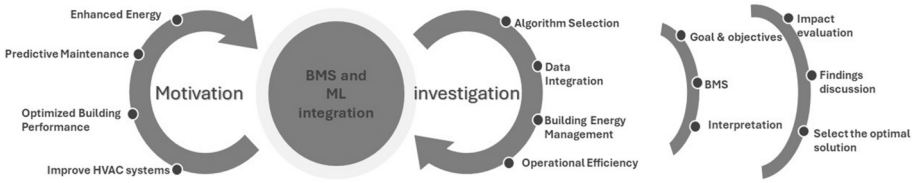

Fig. 3 Applications, motivations, constraints, and ML methods in BMSs that are considered in this review

ity and impact within smart city infrastructures, offering directions for ongoing research to enhance ML-BMS integration for sustainable urban development.

2 A review of machine learning techniques in BMS

The utilization of AI and ML technologies in Building Management Systems (BMSs) is increasingly prevalent, enhancing various aspects of building functionality, including operational efficiency, energy conservation, occupant well-being, and maintenance. This integration relies on data gathered from numerous sensors installed within buildings, enabling the system to make informed decisions and streamline control procedures.These studies highlight how ML models contribute to energy efficiency, predictive maintenance, HVAC optimization, fault detection, and occupancy-based control, demonstrating the transformative impact of AI-driven solutions in smart building. Common ML techniques employed in BMSs include fault detection, energy prediction and optimization, and advanced control strategies. In the following section, we will provide a comprehensive examination of these ML-based methods, outlining their objectives, functionalities, and practical applications. Table 3 summarizes recent studies that utilize various ML techniques in building energy management. This table provides insights into the applications, levels, ML tasks, and algorithms used, demonstrating the breadth of research in this area.

Figure 4 illustrates the various applications of BMSs, showcasing how ML techniques can be integrated into energy management systems. This visual representation complements the discussions in this section, providing a clearer understanding of the different areas where ML can be applied effectively. ML-based techniques in BMS span a wide array of applications beyond those depicted, from real-time monitoring of air quality to predictive maintenance of essential systems, each adding to a building's operational resilience and adaptability. Incorporating these ML-driven strategies within BMS frameworks not only enhances control over environmental variables but also facilitates deeper insights into system performance and potential faults before they impact occupants or energy efficiency. The next section will delve into the specific techniques utilized in energy optimization, maintenance prediction, and anomaly detection, providing a detailed overview of the algorithms and methods that are currently driving advancements in this field.

2.1 Deep learning

In recent years, deep learning techniques have been increasingly applied to predict and optimize energy consumption in buildings. El-Maraghy et al. (2024) developed a CNN model for predicting energy consumption in mosque buildings, achieving a MAPE of 4.5%. This

Table 3 Summary of recent studies utilizing ML techniques in building energy management

References	Year	Application	Level	ML task	ML algorithms
Kim and Cho	2019	Energy consumption	Building	Regression	CNN-LSTM
Somu et al.	2021	Energy consumption	Building	Regression	kCNN-LSTM
El-Maraghy et al.	2024	Energy consumption	Mosque	Regression	CNN
Zhang et al.	2024	Energy consumption	City	Regression	CNN
Feng et al.	2024	HVAC fault diagnosis	HVAC	Classification	Attention-based Transfer Learning
Wu et al.	2024	HVAC fault diagnosis	HVAC	Classification	Composite Neural Network
Wu et al.	2024	Occupancy detection	Building	Classification	CNN
Somu et al.	2021	Thermal comfort prediction	Building	Regression	TL CNN-LSTM
Karaiskos et al.	2024	Indoor air quality	Building	Regression	LSTM-RNN
Wu et al.	2022	Predictive maintenance	Equipment	Regression	LSTM-RNN
Javed et al.	2016	Energy optimization	HVAC	Regression	RNN
Tukymbekov et al.	2021	Street lighting control	Infrastructure	Regression	LSTM
Jeon and Kim	2021	Temperature set-point optimization	HVAC	Regression	LSTM
Jang et al.	2022	Heating energy consumption	Building	Regression	LSTM
Karijadi and Chou	2022	Energy consumption	Building	Regression	RF, LSTM
Durand et al.	2022	Appliance consumption data analysis	Building	Regression	LSTM
Wang et al.	2020	Energy consumption prediction	Building	Regression	LSTM
Luo and Oyedele	2021	Energy consumption forecasting	Building	Regression	Adaptive LSTM optimized by GA
Hu et al.	2023	Predictive maintenance	Equipment	Regression	Parallel LSTM-Autoencoder
Matsukawa et al.	2019	Maintenance operations prediction	Equipment	Regression	LSTM
Zhu et al.	2022	HVAC fault detection	HVAC	Classification	LSTM-SVDD
Patil et al.	2024	Energy performance forecasting	Building	Regression	ANN, RSM
Bhagwat et al.	2024	Fault detection	Infrastructure	Classification	ANN
Ren et al.	2023	Energy efficiency optimization	Building	Regression	ANN
Olanrewaju and Tan	2022	Maintenance satisfaction analysis	Building	Regression	ANN
Abdelaziz et al.	2023	Energy consumption forecasting	Building	Clustering	PCA, SOM, K-means, GA
Arias-Requejo et al.	2023	HVAC control	Building	Clustering	K-means, ICA
Ramírez-Sanz et al.	2023	Fault detection	Equipment	Classification	SSL
Liu and Gou	2024	Indoor thermal comfort control	HVAC	RL	RL
Fährmann et al.	2022	Anomaly detection	Building	Detection	DDQN

(continued) Table 3

References	Year	Application	Level	ML task	ML algorithms
Ding et al.	2022	Multi-zone HVAC control	Building	Deep RL	Deep RL
Fu et al.	2018	Energy consumption	Building	RL	SARSA
Alfaverh et al.	2020	Peak energy demand management	Infrastructure	RL, Fuzzy Reasoning	RL, Fuzzy Reasoning
Geng et al.	2022	Indoor air quality monitoring	Building	Clustering	Clustering
Oliosi et al.	2023	Anomaly detection	Building	Detection	PCA, Spectral Clustering
Wen et al.	2023	Fault detection	Equipment	Detection	PCA
Chen et al.	2023	Temperature and occupancy detection	Building	Classification	SSL
Parhizkar et al.	2021	Energy consumption prediction	Building	Clustering	PCA
Fan et al.	2024	HVAC fault detection	HVAC	Classification	Active Learning, SSL
Nguyen et al.	2021	Real-time energy monitoring	Building	Clustering	Clustering, Regression
Pekşen et al.	2024	Predictive maintenance	Equipment	Classification	SSL
Ahn and Park	2020	HVAC system efficiency	HVAC	DQN	Deep Q-Network
Xu et al.	2021	Fault detection	HVAC	RL	RL
Wei et al.	2020	Energy management	Building	Actor-critic RL	Actor-Critic RL
Jendoubi and Bouffard	2023	Energy management	Building	Optimization	HRL
Qin et al.	2022	Energy optimization	Building	Hybrid RL-GA	RL, Genetic Algorithm
Ji et al.	2019	Energy management	Building	RL	Real-time RL
Quang and Phuong	2024	Energy optimization	Residential HVAC	Deep RL	Deep RL
Zhang et al.	2022	Energy management	Multiple buildings	MARL	Multi-agent RL
Han et al.	2021	Occupant comfort	Building	RL	RL
Brandi et al.	2020	Indoor temperature and energy usage	Building	RL	RL
Masdoua et al.	2023	Fault-tolerant control	HVAC	RL	RL
Fang et al.	2023	Lighting control	Building	RL	RL
Shen et al.	2022	Energy control systems	Building	RL	Multi-agent deep RL

performance is 12% better than an ANN model's MAPE of 5.36% for residential buildings, and it also outperforms other models such as Random Forest (MAPE of 6.023%) and ANN (MAPE of 6%). While CNN models demonstrate strong predictive capability, their performance can be influenced by the complexity of the dataset and model scalability. To enhance predictive accuracy, Kim and Cho (2019) introduced a CNN-LSTM-based model for predicting residential energy consumption that integrates spatial and temporal dependencies. While CNN-LSTM achieves a lower MSE of 0.3738-showing a 49.8% improve-

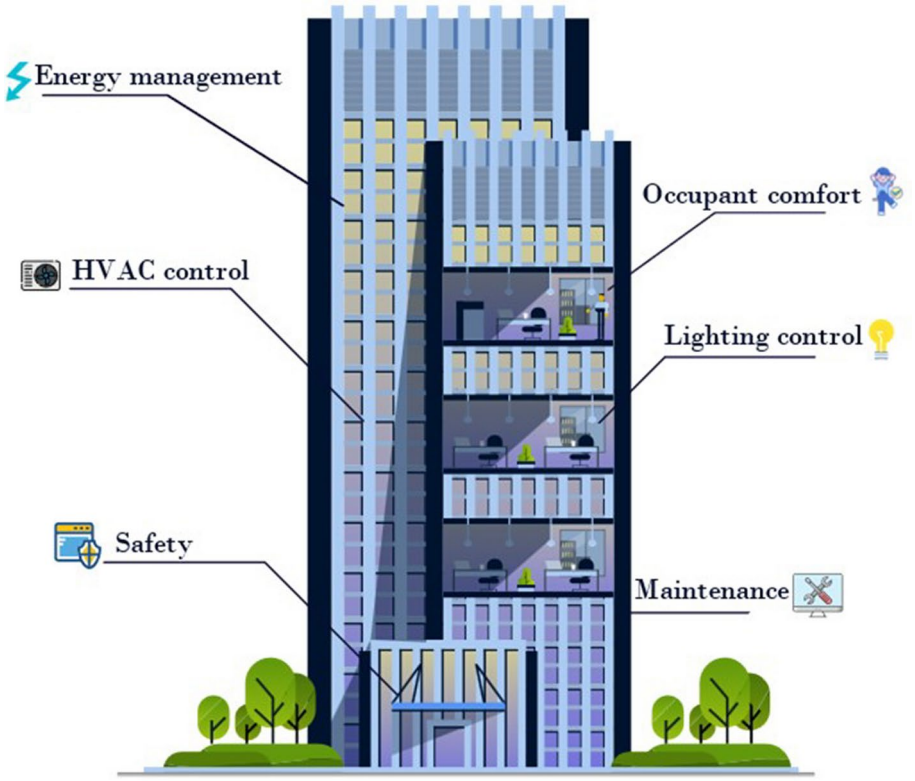

Fig. 4 Illustration of various applications of BMS utilizing ML techniques. This figure highlights key areas such as energy management and predictive maintenance, demonstrating how ML enhances operational efficiency and occupant comfort

ment over Attention LSTM (0.6984)-it tends to be more effective for aggregated time-series data rather than minute-level fluctuations. CNN-LSTM underperforms compared to LSTM and Linear Regression, with approximately 50% lower metrics in minutely data. Nonetheless, it significantly outperforms LSTM, GRU, Bi-LSTM, and Attention LSTM. Further refinements were introduced by Somu et al. (2021), who proposed the kCNN-LSTM model, incorporating k-means clustering to refine input data segmentation. The kCNN-LSTM achieves a MAPE of 0.1670, performing well in weekday and weekend energy forecasting by leveraging structured patterns within the dataset. However, the absence of key contextual factors like occupancy data highlights the ongoing challenge of capturing the full complexity of building environments. To improve the learning of complex temporal dependencies, Wu and Wu (2024) introduced the CNN-BiLSTM-SA model, which combines bidirectional LSTM layers with self-attention mechanisms. This model reduces RMSE by 82.70% compared to BiLSTM and 43.24% compared to BiLSTM-SA. Additionally, it achieves the highest R^2 value among CNN, LSTM, and CNN-LSTM models, demonstrating its effectiveness in accurately predicting household electricity consumption. Its ability to capture both past and future dependencies, along with attention-based feature selection, enhances its predictive performance in dynamic energy consumption scenarios. Although this approach show-

cased promising results, its dependence on small datasets and simulated environments-as also observed in studies such as El-Maraghy et al. (2024) and Zhang et al. (2024c)-raises concerns about scalability and real-world applicability. To overcome these limitations, researchers have explored hybrid frameworks that integrate clustering and transfer learning strategies.

LSTM networks, in particular, have shown considerable promise in capturing temporal trends critical to energy systems management. Enhanced interpretability, achieved through techniques like layer-wise relevance propagation in LSTM models (Wu et al. 2022), has provided new insights into predictive maintenance strategies. Moreover, the integration of random neural networks in cloud-enabled smart controllers (Javed et al. 2016) has yielded substantial energy savings of 27.12% compared to traditional rule-based systems. A variety of studies (Tukymbekov et al. 2021; Jeon and Kim 2021; Jang et al. 2022; Karijadi and Chou 2022; Durand et al. 2022; Wang et al. 2020; Luo and Oyedele 2021; Hu et al. 2023; Matsukawa et al. 2019; Zhu et al. 2022) further attest to the versatility of LSTM networks-from optimizing street lighting based on weather forecasts to real-time HVAC fault detection and indoor air quality prediction-demonstrating both their potential and the necessity for further refinement. For HVAC systems, attention-based transfer learning methods (Feng et al. 2024) and composite neural network approaches (Wu et al. 2024a) have been deployed to address sensor fault diagnosis and data imbalance, respectively. These methodologies highlight the critical role of deep learning in fault detection and diagnosis within BMSs. Beyond LSTM networks, artificial neural networks (ANNs) have also been extensively employed for energy management tasks. Simulation-based ANN frameworks (Roodkoly et al. 2024) and models integrating ANN with Response Surface Methodology (Patil et al. 2024) have proven effective in forecasting energy metrics and optimizing building performance. Furthermore, ANN-driven fault detection systems (Bhagwat et al. 2024) and intelligent control frameworks for public buildings (Ren et al. 2023) illustrate the capacity of these networks to enhance operational efficiency and occupant comfort, even in contexts demanding post-pandemic adjustments (Olanrewaju and Tan 2022). Collectively, these studies reveal a dynamic landscape in which deep learning methods are continuously evolving to address the multifaceted challenges of building energy management. Although each approach-whether CNN-based, LSTM-centric, or ANN-driven-offers unique advantages, common challenges persist. These include the need for larger, more diverse real-world datasets, improved model interpretability, and the integration of multi-modal sensor data. Future research should aim to develop hybrid models that strike a balance between predictive accuracy and practical applicability, ultimately bridging the gap between simulation and real-world deployment. Cordeiro-Costas et al. (2024) propose a hybrid methodology combining LSTM and Multilayer Perceptron (MLP) models, optimized with the Non-dominated Sorting Genetic Algorithm II (NSGA-II). This method uses Global Forecast System (GFS) data to predict energy consumption and optimize distributed energy sources like photovoltaic (PV) systems. By balancing energy costs and efficiency, NSGA-II identifies optimal solutions along the Pareto front. Implemented at the Industrial Engineering School of the Universidade de Vigo, Spain, this approach effectively enhances hyperparameter tuning and energy balance, showcasing the potential of integrating machine learning with optimization for better energy management in buildings.

As illustrated in Fig. 5, various machine learning approaches, including LSTMs, ANNs, and CNNs, play distinct yet interrelated roles in building energy management. These meth-

ods are applied across domains such as energy efficiency, HVAC optimization, occupancy detection, and predictive maintenance, underscoring their versatility in creating adaptive and efficient building environments. This integration of advanced ML algorithms in BMSs not only streamlines operations but also fosters sustainability by optimizing resource use and enhancing occupant comfort. The diversity of approaches highlighted in recent studies demonstrates the adaptability and potential for future innovations within smart building systems, setting the stage for increasingly resilient, responsive, and sustainable building management solutions.

2.2 Supervised learning

Supervised learning is a cornerstone methodology in analyzing annotated energy datasets, demonstrating robust performance across various applications in BMSs. However, the reliance on labeled data presents significant challenges for real-world deployment. Obtaining

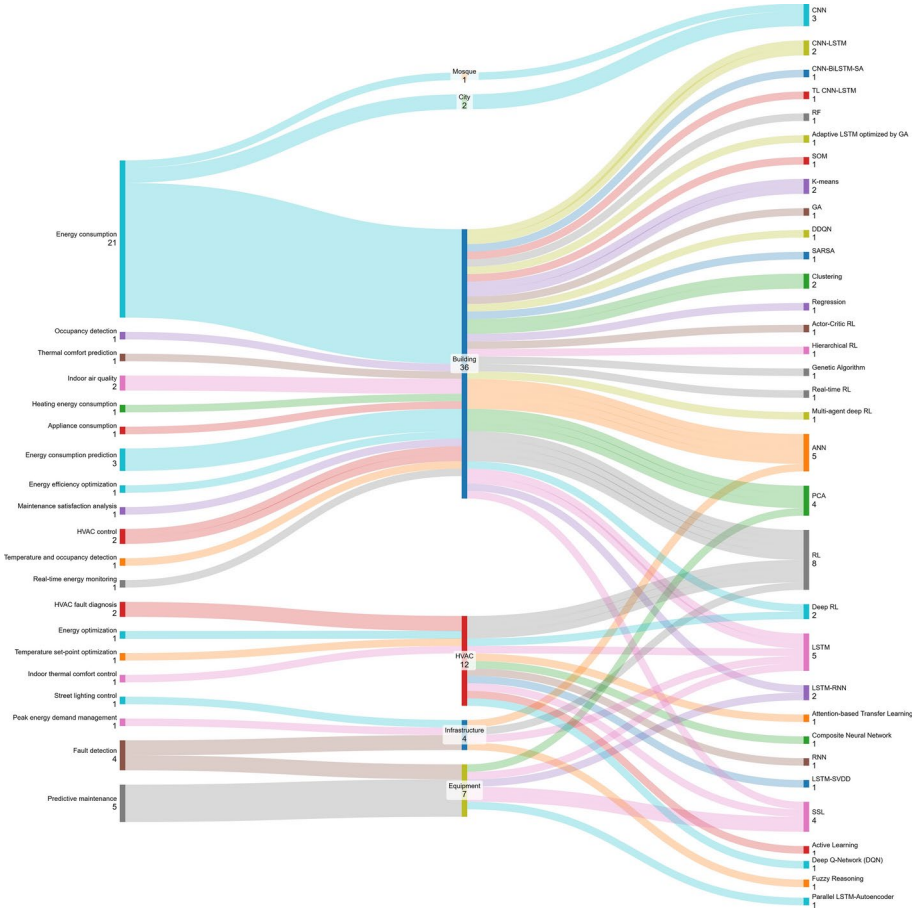

Fig. 5 Sankey diagram illustrating the applications of ML and deep learning in BMS. This figure highlights how various ML techniques are integrated across different domains, including buildings, cities, and infrastructure

high-quality labeled datasets is not only time-consuming but also costly, which can hinder the scalability of supervised learning techniques in practical energy applications. This limitation is particularly pronounced in sectors where data labeling is scarce or expensive, potentially stifling innovation and efficiency improvements in energy management.

Classification models: Classification models are traditional yet powerful tools for tasks such as energy prediction, indoor activity monitoring, and fault detection. Key algorithms, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Trees (DT), have shown varying degrees of effectiveness. For instance, Zhang et al. (2022b) achieved an impressive accuracy rate of 99.98% using a hybrid PSO-SVM algorithm for energy balancing in green buildings. Such results not only highlight the potential for enhanced energy efficiency but also suggest a model that can be scaled for broader applications in similar contexts. However, these models have inherent limitations. SVMs, while effective, struggle with non-linear problems unless kernel methods are applied, which adds complexity to their implementation. KNN's performance is heavily dependent on the choice of "K," and the algorithm can become computationally expensive with large datasets. Moreover, Decision Trees can overfit on small datasets, leading to poor generalization-an issue that can significantly impact predictive maintenance strategies in operational settings.

Regression models: Regression models, focusing on identifying relationships between variables, play a critical role in energy forecasting and anomaly detection. Common methods, including Support Vector Regression (SVR) and Random Forest (RF), have been widely adopted. Notably, Moulla et al. (2024) utilized a diverse dataset to predict hourly energy consumption, achieving high accuracy with RF and DT models. Their findings offer valuable insights that can inform energy management policies and practices, particularly in regions grappling with load-shedding crises. These widely adopted machine learning models exhibit ease of implementation and interpretability. However, the assumption of linearity between variables may not always hold true in complex real-world scenarios. Addressing these challenges through advanced hybrid models could enhance predictive accuracy and operational efficiency in BMS.

2.3 Unsupervised learning

Unsupervised learning techniques have proven versatile and effective in BMSs, significantly enhancing energy efficiency, reducing costs, and improving system robustness. These methods allow for the exploration of data without the constraints of labeled datasets, making them particularly valuable in dynamic environments where data is abundant but unannotated.

Clustering: Clustering algorithms are employed to categorize and forecast energy usage patterns, thereby improving energy management in buildings. Abdelaziz et al. (2023, 2024) developed a comprehensive framework using Principal Component Analysis (PCA) for dimensionality reduction, Self-Organizing Maps (SOM) for pattern identification, and K-means clustering combined with a Genetic Algorithm (GA) to optimize energy consumption clusters. This innovative methodology not only improved energy demand forecasting but also enhanced load management strategies, effectively reducing energy wastage and operational costs. Similarly, Arias-Requejo et al. (2023) applied K-means clustering and Independent Component Analysis (ICA) for energy consumption forecasting in smart buildings, focusing on HVAC controls and energy-saving strategies. They emphasized the criti-

cal importance of data preprocessing and feature extraction to handle correlations among variables, which can significantly impact forecasting accuracy. Raja and Saraswathi (2023) introduced an IoT system leveraging hierarchical clustering and Gaussian Mixture Models (GMM) to classify energy use behaviors and model occupancy patterns, demonstrating high accuracy and minimizing overall power consumption. Additionally, Tian et al. (2024) utilized association rule mining alongside clustering and rule-based methods to optimize energy consumption and enhance fault detection in HVAC systems. Their unsupervised data mining-based framework resulted in a 6.9% energy savings, underscoring the practical benefits of applying clustering techniques in real-world settings. Furthermore, Wang et al. (2022a) and Etezadifar et al. (2023) investigated clustering approaches for event-based non-intrusive load monitoring (NILM) and appliance identification, demonstrating significant contributions to energy performance evaluation and ranking in workplaces. Liu et al. (2018) focused on energy efficiency assessment in industrial buildings, while Gunay and Shi (2020) applied clustering to detect operational anomalies in building automation systems. The exploration of indoor air quality (IAQ) monitoring, as examined by Sha et al. (2023) and Geng et al. (2022), further establishes the broad applicability of clustering techniques in optimizing building performance.

Dimensionality reduction: Dimensionality reduction techniques are crucial for simplifying complex datasets, enhancing the efficiency of anomaly detection, and improving energy management systems. Abdelaziz et al. (2023, 2024) used PCA for dimensionality reduction, which contributed significantly to their framework for optimizing energy consumption clusters. By reducing the feature space, PCA aids in identifying key variables that drive energy usage, streamlining subsequent modeling efforts. Oliosi et al. (2023) implemented PCA and spectral clustering to reduce the dimensionality of complex sensor data, improving the efficiency of anomaly detection and maintenance. Wen et al. (2023) applied PCA for early fault detection and classification, while Parhizkar et al. (2021) and Baird et al. (2017) demonstrated its utility in energy consumption prediction and occupancy detection, respectively. Khan et al. (2020) applied the t-Stochastic Neighborhood Embedding (t-SNE) algorithm to eliminate redundant features, thus preventing large coefficients and improving model performance. In a novel approach, Miyasawa et al. (2019) introduced an energy breakdown technique using smart meter data and semi-binary nonnegative matrix factorization (SBNMF) to estimate individual appliance power consumption without additional sensors. To enhance SBNMF accuracy, the authors proposed three model assumptions and developed appliance-level classifiers using random forest, incorporating auxiliary information like user feedback to improve performance. Song et al. (2022) compared six machine learning algorithms and found Linear Discriminant Analysis (LDA) to be more accurate for thermal comfort evaluation, and Lee et al. (2020) used LDA for building flow detection, highlighting its broad applicability in BMSs. These studies collectively underscore the importance of unsupervised learning techniques in enhancing the efficiency and effectiveness of BMSs, particularly through clustering and dimensionality reduction methodologies.

2.4 Semi-supervised learning

Semi-supervised learning (SSL) is a crucial methodology in the development of BMSs, particularly in scenarios where obtaining comprehensive labeled datasets is difficult or expensive. By leveraging both annotated and unannotated data, SSL techniques can significantly

improve model performance and predictive accuracy, addressing some of the limitations inherent in fully supervised methods. SSL has diverse applications within BMS, particularly in automated fault detection and diagnosis (AFDD) for HVAC systems. Dey et al. (2018) underscore its utility in managing unstructured and unlabeled HVAC sensor data, enhancing system reliability and reducing the operational costs associated with data preprocessing. In building energy modeling (BEM), Naganathan et al. (2016) explored the use of clustering algorithms and semi-supervised machine learning to optimize energy efficiency by analyzing real-time data from substations and buildings. This approach not only identifies factors contributing to energy losses but also aids utility providers in effective energy supply–demand management. Akbar et al. (2024) introduced an innovative SSL-based deep learning framework for non-intrusive load monitoring (NILM) in smart grids, disaggregating aggregate energy consumption data into individual appliance-level insights. This methodology enhances energy optimization and cost reduction, demonstrating superior accuracy compared to traditional methods. The integration of active learning with SSL further enhances data-driven HVAC fault diagnosis, reducing labeling costs and improving system reliability, as noted by Fan et al. (2024). This combined approach effectively identifies valuable data for fault detection, supporting practical applications in real-world settings. Additionally, Ramírez-Sanz et al. (2023) provide a comprehensive review of SSL applications in industrial fault detection and diagnosis, highlighting its effectiveness in handling limited labeled data and improving model accuracy across various industrial environments. SSL's versatility in BMS extends to energy consumption forecasting, predictive maintenance, IoT integration, and renewable energy utilization. For instance, Chen et al. (2023b) exemplify its use in accurate indoor temperature prediction and occupancy detection by leveraging both labeled and unlabeled sensor data. Hybrid SSL models that combine clustering and regression approaches have demonstrated promise in real-time energy monitoring and predictive maintenance within BMS, thereby enhancing operational efficiency and prediction accuracy, as shown by Nguyen et al. (2021) and Pekşen et al. (2024). Despite its advantages, SSL can exhibit instability in results and lower performance compared to fully supervised learning when labeled data is insufficient, as highlighted by Wang et al. (2022b). Nevertheless, SSL's ability to effectively utilize both labeled and unlabeled data positions it as a valuable tool for addressing data scarcity challenges and enhancing overall BMS performance.

2.5 Reinforcement learning

Reinforcement learning (RL) has shown significant promise in optimizing various aspects of BMSs, including HVAC control, maintenance, fault detection, energy prediction, and consumption. A study by Ahn and Park (2020) explored the use of Deep Q-Networks (DQN) to enhance HVAC system efficiency and occupant comfort. This approach illustrates how RL can dynamically adjust system parameters to achieve optimal performance. Similarly, Liu and Gou (2024) introduced an RL model that improved indoor thermal comfort by 24% and reduced air conditioning usage by 24.7% compared to baseline models. These results not only highlight the effectiveness of RL in energy savings but also demonstrate its potential impact on occupant satisfaction. Fährmann et al. (2022) employed deep Q-learning (DDQN) for anomaly detection in smart buildings, showcasing RL's adaptability in identifying and responding to unusual behaviors in energy consumption. In the realm of energy-

efficient control, Xu et al. (2021) applied RL to fault detection and diagnostics in HVAC systems, ensuring efficient operation. Ding et al. (2022) proposed a deep RL-based method for controlling thermal comfort in multi-zone residential HVAC systems. Fu et al. (2018) utilized SARSA to predict and minimize energy consumption in commercial buildings. Wei et al. (2020) implemented actor-critic methods for smart building energy management, demonstrating improved efficiency. The versatility of RL is further exemplified by Jendoubi and Bouffard (2023), who applied hierarchical RL for managing complex building energy tasks, enhancing system responsiveness. Qin et al. (2022) combined genetic algorithms with RL for enhanced energy optimization, while Ji et al. (2019) explored real-time RL for energy management in smart buildings. Moreover, Quang and Phuong (2024) developed a deep RL algorithm to optimize energy consumption in residential HVAC systems while maintaining occupant comfort. Zhang et al. (2022a) employed multi-agent RL for coordinated energy management across multiple buildings, and Alfaverh et al. (2020) applied RL and fuzzy reasoning to manage and reduce energy demand during peak periods. Han et al. (2021) balanced occupant comfort using RL, and Brandi et al. (2020) enhanced indoor temperature and energy usage with RL. Masdoua et al. (2023) developed fault-tolerant HVAC control strategies with RL, while Fang et al. (2023) optimized lighting control systems using RL for energy savings. In a significant advancement, Shen et al. (2022) introduced a multi-agent deep RL optimization framework for building energy systems incorporating renewable energy, utilizing a dueling double deep Q-network for single-agent optimization and a value-decomposition network for multi-agent cooperation. These studies collectively highlight the vast potential of reinforcement learning to enhance efficiency, reduce costs, and improve occupant comfort in BMSs.

2.6 Generative-AI, federated learning, and transfer learning

The rapid evolution of AI, exemplified by advanced language models like ChatGPT, holds significant promise for specialized engineering tasks, particularly in physics-based building energy modeling (BEM) (Zhang et al. 2024a). These models simplify data analysis and generate simulation inputs, demonstrating their utility in enhancing modeling processes. However, their effectiveness depends on selecting appropriate techniques, such as prompt engineering or integration within multi-agent systems. Despite challenges like computational demands and self-consistency issues, advancements are expanding the use of language models across various sectors (Alqahtani et al. 2023). In the context of Federated Learning (FL), a decentralized approach allows models to be trained across multiple devices while preserving data privacy (Li et al. 2021). FL encompasses horizontal federated learning (HFL), which aggregates models from devices with similar features but different samples, and vertical federated learning (VFL), which integrates diverse feature sets from the same samples. Applications in smart buildings demonstrate FL's effectiveness in anomaly detection and thermal comfort management, showcasing its potential to enhance operational efficiency and user satisfaction (Sater and Hamza 2021; Khalil et al. 2021). HFL and VFL enhance model accuracy and efficiency while addressing privacy concerns (Wang and et al. 2023; Liu et al. 2024). Transfer Learning (TL) enables the reuse of models trained on one task for related tasks, particularly when data is scarce. TL methods include pre-training on large datasets followed by fine-tuning on smaller, task-specific datasets. Applications range from intelligent fault diagnosis to energy demand forecasting, showcasing significant

improvements in accuracy and performance, thus facilitating more informed decision-making in energy management (Chen et al. 2023a; Coraci et al. 2023). In summary, the integration of generative AI, FL, and TL presents unprecedented opportunities for enhancing knowledge management, decision-making, and innovation in engineering and construction, leading to more efficient, safe, and sustainable practices.

2.7 Fast machine learning in building management systems

ML has become essential across numerous industries, including healthcare, finance, and autonomous vehicles. As the complexity of these applications increases, there is a growing demand for faster and more efficient ML techniques, collectively referred to as FastML. FastML addresses this need by accelerating various stages of the ML pipeline, from data preprocessing to model training and inference. This approach is particularly vital for scenarios requiring real-time or near-real-time decision-making, such as scientific research, BMSs, and autonomous vehicles (Duarte et al. 2022b). By reducing the time and computational resources needed for model training, FastML facilitates quicker innovation cycles and deployment of ML solutions. It aims to overcome challenges inherent in traditional ML approaches, including rising computational requirements for training and inference, the need for low latency in certain applications, concerns about energy consumption, and the difficulty of scaling traditional ML methods to handle large-scale datasets and complex models (L'heureux et al. 2017). Key techniques in FastML include leveraging specialized hardware, optimizing models through pruning and quantization, developing efficient neural network architectures, utilizing distributed computing frameworks, and improving algorithms. These techniques enable a range of applications, from real-time computer vision and natural language processing to personalized recommendations and rapid decision-making in financial services and healthcare. FastML refers to techniques designed to accelerate the machine learning pipeline, which includes data preprocessing, model training, and inference (Deiana et al. 2022b). These techniques utilize specialized hardware (e.g., FPGAs, ASICs, or GPUs), optimized algorithms, and advanced strategies to facilitate real-time or near-real-time decision-making. In the context of BMSs, "fast" typically denotes systems capable of processing sensor data and generating control decisions within milliseconds, allowing for rapid responses to dynamic environmental changes. FastML is particularly beneficial in BMS for tasks such as fault detection, energy efficiency optimization, and predictive maintenance, where low latency and high computational efficiency are essential. However, achieving "fast" performance often necessitates careful resource management, especially when deploying complex algorithms like Nonlinear Model Predictive Control (NNMPC) on hardware platforms such as FPGAs. In the study by Fan et al. (2022), the FastML framework was developed to address predictive uncertainty, model drift, and unexpected conditions using Bayesian neural networks (BayesCNNs). These networks provide probabilistic predictions and quantify uncertainty by estimating prediction distributions, enabling nuanced outcomes. Techniques like Monte Carlo dropout enhance uncertainty assessment, making FastML valuable in dynamic environments such as severe weather, building renovations, or COVID-19 lockdowns (Melosik et al. 2022). The framework incorporates adaptive decision-making, adjusting predictions and confidence intervals during uncertainty, and employs continuous learning to address model drift. Key strategies include Bayesian inference, adaptive learning, and probabilistic output handling for risk-aware decisions. FastML

achieves remarkable performance, including up to 92x higher energy efficiency than CPUs, 76x higher than GPUs, 99.39% accuracy, and 9-30x higher throughput than existing accelerators, making it a reliable tool for mission-critical tasks like energy management, fault detection, and predictive maintenance. A cutting-edge framework designed to facilitate the deployment of machine learning models on hardware accelerators, particularly Field-Programmable Gate Arrays (FPGAs) and Application-Specific Integrated Circuits (ASICs), is High-Level Synthesis for Machine Learning (HLS4ML). Unlike traditional software-based implementations, HLS4ML utilizes high-level synthesis (HLS) tools to convert high-level model descriptions into hardware logic, enabling efficient and real-time inference on these platforms. The HLS4ML toolchain transforms these optimized ML models into hardware specifications suitable for implementation on FPGAs and ASICs. This framework not only enhances the performance of machine learning applications but also significantly reduces the time required for deployment. By utilizing hardware accelerators, HLS4ML addresses the need for low-latency inference, making it particularly advantageous for applications that demand rapid decision-making. The integration of hardware and software through HLS4ML paves the way for more efficient and scalable machine learning solutions across various industries, particularly through its incorporation of several core features that are critical for optimizing ML models for hardware deployment:

- **Quantization**: Fixed-point quantization reduces model complexity and resource usage by implementing activation functions and mathematical operations using fixed-point arithmetic, thereby enhancing computational efficiency.
- **Framework support**: HLS4ML integrates with several prominent machine learning frameworks, including (Q)Keras, PyTorch, and QONNX, allowing users to convert models into an internal representation (`HLSModel`) for further optimization and hardware synthesis.
- **Optimization**: The framework performs optimization passes to merge compatible layers, reducing latency and improving hardware resource utilization. Collaboration with the FINN team on QONNX enhances support for quantized neural network models, ensuring smooth interoperability with various back-end tools.

Originally developed for high-energy physics applications, HLS4ML has since expanded to meet the demands of fields requiring low-latency, high-throughput, and energy-efficient inference. It primarily supports Vivado HLS for Xilinx FPGAs but also offers back-ends for Intel HLS and experimental support for Vitis HLS. A key enabler of FastML is the use of specialized hardware components such as GPUs and FPGA. For critical BMS applications like fault detection and energy management, FPGAs have demonstrated superior performance compared to GPUs. Recent benchmarks indicate that FPGAs can reach speeds that are up to 36 times faster and provide energy efficiency improvements of up to 21 times, all while ensuring an optimal balance between latency and throughput (Guo 2024). This makes FPGAs particularly suitable for real-time applications in dynamic building environments. To further enhance performance, FastML employs various optimization techniques. One such technique is feature selection, which reduces the number of input types and focuses on essential data, streamlining the preprocessing stage and improving efficiency without compromising prediction accuracy (Zhang et al. 2024b). Another important approach is data reduction, which involves utilizing fewer meteorological data types or other input variables,

significantly reducing computational overhead while maintaining model performance. Additionally, model quantization plays a critical role by reducing the precision of model parameters, enabling faster inference and lower energy consumption. This is particularly beneficial in resource-constrained environments where hardware limitations necessitate lightweight models (Rutishauser 2024). Despite its advantages, FastML faces challenges in BMS applications. For instance, while techniques like quantization and data reduction enhance efficiency, they may also introduce trade-offs in model accuracy or robustness. Additionally, the deployment of FastML models on hardware like FPGAs often requires specialized tools and expertise, which can hinder developer productivity and slow down implementation. As noted by Wang et al. (2019), the complexity of hardware programming presents a significant barrier. Furthermore, while FPGAs offer computational acceleration through reconfigurability, the time required for reconfiguration can be a significant drawback. To address these challenges, this paper utilizes HLS4ML, a tool designed to convert machine learning models into hardware designs (FastML Team 2024). HLS4ML strikes a balance between hardware efficiency-ensuring optimal performance on devices like FPGAs-and developer productivity-simplifying the process of creating and deploying AI applications. By reducing the time required for FPGA deployment, HLS4ML enables faster and more accessible implementation of FastML solutions in BMS. This approach not only enhances the performance of critical applications like fault detection and energy management but also supports the broader adoption of FastML in building management. FastML is revolutionizing building management by leveraging advanced algorithms to analyze real-time data from IoT devices. This approach facilitates proactive adjustments in operations, enhancing energy efficiency and supporting sustainability initiatives. Barbaresi et al. (2022) demonstrate the effectiveness of various machine learning models, particularly tree-based methods, in predicting building energy requirements and improving design strategies. Seyedzadeh et al. (2019) stress the need for model fine-tuning to achieve precise heating and cooling load predictions, which are crucial for optimizing energy consumption. Beyond energy prediction, FastML contributes to the development of advanced energy materials, as highlighted by Farhadi et al. (2023), promoting further improvements in energy performance. Deiana et al. (2022a) discuss the integration of machine learning into scientific processes, offering valuable insights for BMSs. Aarrestad et al. (2021) explore the deployment of low-latency neural networks on field-programmable gate arrays (FPGAs) for real-time data processing in building management, while Ngadiuba et al. (2020) focus on optimizing resource use within machine learning applications, enhancing BMS efficiency. Additionally, the review by Seyedzadeh et al. (2018) outlines various machine learning methodologies that enhance building energy performance, laying the groundwork for adaptive operational frameworks in smart buildings. Dey et al. (2020) proposed a machine learning-based multi-level framework designed to enhance functionality and enable quick fault detection in smart buildings. However, this approach is computationally expensive due to a lack of model optimization, particularly when applied to large-scale buildings or in real-time scenarios. The high computational demands pose significant challenges, especially for older BMSs with limited processing power, which may hinder the feasibility of real-time fault detection and response. To address these issues and accelerate machine learning while reducing complexity, optimization techniques such as quantization are essential. This optimization would facilitate deployment on resource-constrained devices by leveraging FPGA acceleration. To further tackle these challenges, Agouzoul et al. (2022) developed an efficient energy man-

agement system that utilizes a Model Predictive Control (MPC)-based ANN implemented on FPGA technology. Their simulation-based approach enhances real-time processing capabilities by enabling the parallel execution of control algorithms. This innovation not only improves energy consumption optimization but also reduces computational overhead and increases processing speed. However, it introduces programming complexities, as implementing such systems requires specialized skills in hardware description languages (HDL) like VHDL and Verilog, along with meticulous resource management. Deploying intricate algorithms such as NNMPC on an FPGA necessitates careful planning of resources, including logic elements and memory blocks, to ensure efficient hardware utilization and optimal performance. Additionally, Sen et al. (2023) proposed a fast, machine learning-based predictive control approach for energy management systems (EMS). Their study emphasizes the importance of real-time data for effective implementation while also acknowledging the potential risks associated with the reprogrammability of FPGA systems.In summary, FastML significantly enhances the capabilities of BMS in areas such as predictive maintenance, occupancy detection, and energy forecasting, all of which are crucial for developing smart, sustainable buildings. However, to fully realize the potential of FastML, challenges related to data quality and model interpretability must be addressed. Future research should explore the integration of FastML with smart grid technologies to enable dynamic responses to fluctuations in energy supply and demand. Collaborative efforts among experts in engineering, data science, and environmental science will be vital in overcoming the complex challenges faced in modern building management.

3 Evaluation and results

In this section we present how FastML and HLS4ML can be applied in an building management energy forecasting application. This section uses data and models form a real case study (Queen's Building) and provides the following contributions (i) a forecasting capability to prediction energy consumption in an educational building context and (ii) a fast machine learning capability using high level specification ML that demonstrates the use of FastML for BMSs.

3.1 Experimental testbed and pilot

The evaluation part describes an energy consumption prediction for optimizing energy in buildings. The Queen Building, depicted in Fig. 6, serves as a case study for this analysis, showcasing its architectural layout. The objectives of this evaluation are to (i) develop a forecasting model using traditional machine learning techniques and (ii) transform the ML models using HLS4ML to demonstrate effectiveness in energy forecasting and energy management based on real data and models from a case study example. LSTM models, a type of Recurrent Neural Network (RNN), were selected for their proven effectiveness in capturing long-range dependencies in sequential data, which is critical for accurately modelling time-series patterns in energy consumption. Energy usage in buildings often exhibits temporal trends, such as daily or weekly cycles, and LSTMs are particularly well-suited for capturing these dynamics. Additionally, their compatibility with the HLS4ML framework, which enables efficient deployment on resource-constrained hardware, was a key factor in our

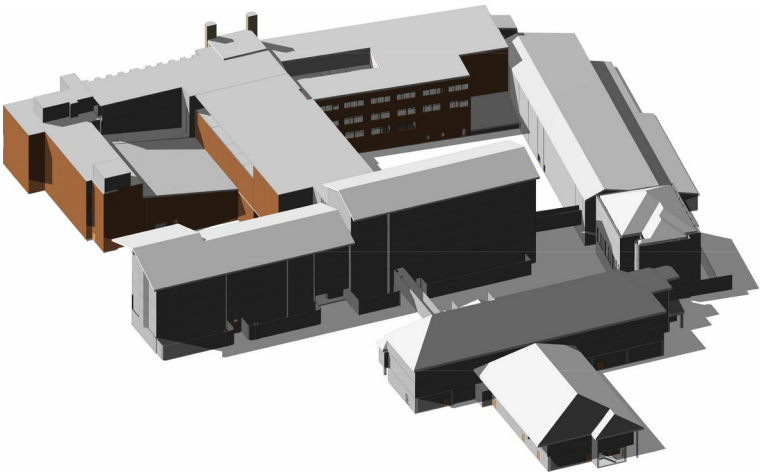

Fig. 6 Architectural layout of the Queen's Building, the primary case study for evaluating energy consumption prediction models. This figure illustrates the framework for preparing the LSTM model, including data extraction, validation, and the training process on an HLS-based FPGA

model selection. While alternative architectures, such as Transformers and Graph Neural Networks (GNNs), show promise, their computational demands and limited support within HLS4ML rendered them less practical for our current work. However, the computational intensity of LSTMs poses challenges for deployment in resource-constrained environments, such as edge devices. This study investigates four optimization strategies for LSTM models aimed at enhancing energy consumption forecasting: standard LSTM, pruned LSTM, quantized LSTM, and a hardware-accelerated model using HLS4ML. We focus on pruning and quantization techniques that improve model efficiency by reducing computational complexity while maintaining accuracy. The models are evaluated based on three criteria: accuracy, inference speed, and practical feasibility for real-time applications. Although both pruning and quantization effectively reduce model complexity and enhance inference speed, they may slightly compromise accuracy. In contrast, hardware acceleration through HLS4ML can provide substantial performance improvements, though careful integration is necessary to meet hardware constraints. The experimental setup utilized a system running Ubuntu 18.04.2 LTS on an x86 64 architecture, powered by a 13th Gen Intel(R) Core(TM) i7-1355U processor with 12 logical CPUs across 6 cores and hyper-threading. This configuration, featuring a base clock speed of approximately 2.6 GHz and 7.6 GB of RAM, supports efficient multi-tasking and high-speed connectivity via a 10 Gbit/s network interface. Key software tools used in this study include Pandas, NumPy, Seaborn, TensorFlow Model Optimization, HLS4ML, and Vivado for high-level synthesis.

Understanding the architectural layout of the Queen Building is crucial, as it impacts energy consumption patterns. The diverse spaces within the building, including offices, laboratories, and common areas, present unique challenges for energy forecasting and optimization. This study aims to leverage advanced machine learning techniques to provide insights that can inform energy management strategies, ultimately leading to more sustainable practices in building operations.

3.1.1 Dataset

This project investigates the potential of creating an LSTM model to predict energy consumption using a dataset sourced from the Queen Building. The data is formatted as a CSV file and contains 1,044 observations across 21 variables, with no missing or duplicate entries. Load volume statistics reveal a minimum of 7.78 MWh, a maximum of 29.0 MWh, and an average load volume of approximately 16.46 MWh. Key variables include energy usage data for multiple buildings, such as Queens East, Queens North, and others, along with total energy metrics. One critical aspect to consider is that the dataset may contain gaps in timestamps, and zero values do not always indicate no energy consumption; they can represent instances where the meter failed to report. Additionally, to accurately calculate total energy usage for the Trevithick building, adjustments must be made by subtracting excluded energy metrics. These considerations highlight the importance of preprocessing and refining the dataset to ensure accurate predictions. Overall, this project seeks to leverage this detailed dataset to enhance energy consumption forecasting and improve the model using various methods, including HLS4ML. In this context, FastML offers several advantages. First, it allows for seamless compatibility with various BMS architectures, enabling efficient data exchange and control across different platforms. By processing and analyzing large datasets in real time, FastML enhances decision-making, optimizing energy use and occupant comfort regardless of the building type. Furthermore, FastML provides scalable machine learning models that can be tailored to specific building needs, offering customized solutions that improve overall operational efficiency. Moreover, FastML leverages machine learning algorithms that can quickly learn and adjust to diverse regional climates, construction practices, occupant cultural preferences, and economic conditions. For instance, its ability to process large datasets enables it to identify patterns and optimize performance based on specific local factors, such as temperature variations, building materials, and user behaviors. To illustrate, energy consumption patterns in tropical climates differ significantly from those in temperate or arid regions, while cultural norms around comfort and energy use can influence system design and adoption. This flexibility ensures that energy management strategies are tailored to meet the unique needs of different regions, enhancing efficiency and occupant satisfaction while promoting sustainable energy use across various contexts. Finally, the use of standardized APIs further simplifies integration, ensuring that diverse systems can communicate effectively while maintaining data integrity and security. These capabilities ultimately lead to enhanced sustainability and performance across a portfolio of buildings.

3.1.2 Data preprocessing

To prepare the energy consumption data for modeling, several preprocessing steps were carefully designed to address common data quality issues, handle missing values, and engineer useful features to improve model accuracy:

- **Handling missing data:** The dataset contained instances where energy values were recorded as zero, indicative of meter failures rather than actual consumption. Interpreting these zeros as legitimate data could lead to misleading conclusions. To address this,

all zero values in the energy consumption columns were replaced with NaN values. A forward-filling technique was then applied to fill these gaps, ensuring data continuity by using the last valid observation to maintain the integrity of the time series.

- **Resampling:** The original dataset lacked a consistent time interval due to gaps in meter reporting or irregular data logging. To standardize the dataset, it was resampled to 15-minute intervals, calculating the mean of all values within each interval. Remaining gaps were again forward-filled to produce a complete and uniform time series.
- **Feature engineering:** Several temporal features were extracted, including month, year, week number, and day of the week, to capture seasonal consumption patterns. A new feature, "season," categorized the data into Spring, Summer, Autumn, or Winter to help the model capture broader consumption trends influenced by weather or time of year. Additionally, lag features representing energy consumption from the past seven days were introduced, enabling the model to utilize historical behavior in predictions.
- **Special case of Trevithick Data:** The energy consumption data for the Trevithick building required adjustment, as the reported values included data from another building, inflating the figures. To obtain the actual energy consumption for Trevithick, values in the 'Trev' column were adjusted by subtracting the corresponding values in the 'Trev_Exclude' column, ensuring accurate data specific to the building's consumption.
- **Incorporating Domain-Specific Insights:** The preprocessing steps were informed by a deep understanding of energy consumption dynamics. Recognizing that zero values often stemmed from meter failures allowed for more accurate data handling. Extracting seasonal features accounted for predictable variations in energy usage, while lag features utilized the temporal dependencies inherent in energy consumption patterns. Adjusting the pilot data demonstrated awareness of the complexities involved in accurately reporting energy usage across multiple buildings.
- **Scaling:** After preprocessing, the dataset was normalized using the Min-Max scaling method, which transformed values to a range between 0 and 1. This normalization ensured that all features contributed equally during model training, preventing features with larger ranges from dominating the outcomes.

The scaling process involved several steps:

- **Initialization:** The MinMaxScaler was applied to normalize the feature set.
- **Fitting the Scaler:** The scaler was fitted to compute the minimum and maximum values for each feature.
- **Transforming the Data:** The data was normalized using the formula:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

- **Reshaping for LSTM:** The scaled data were reshaped to meet the input requirements of LSTM models, which typically expect three-dimensional input in the format of (samples, time steps, features).

3.1.3 Outlier handling in data preprocessing

The analysis of the "Actual Trev Energy Usage" data revealed significant outliers, particularly energy consumption levels exceeding 100 megawatts (MW). These extreme values could distort model performance and were identified for removal.

The identification and removal of outliers were conducted as follows:

- **Identification of Outliers:** The Interquartile Range (IQR) method was employed to detect outliers. Here, the first quartile (Q1) and the third quartile (Q3) were calculated, yielding the IQR as $\text{IQR} = Q3 - Q1$. Outliers were defined as values falling below $Q1 - 1.5 \times \text{IQR}$ or above $Q3 + 1.5 \times \text{IQR}$.
- **Filtering the Dataset:** The dataset was filtered to retain only data points within acceptable limits, including a check to ensure that maximum values across relevant energy columns did not exceed the threshold of 100 MW. This process involved removing missing values and ensuring that *Actual Trev* values remained positive.
- **Resulting Dataset:** The removal of outliers resulted in a dataset that more accurately represents typical energy consumption patterns, thus reducing potential skew in analyses and visualizations.

3.1.4 Data visualization

Data visualizations were created to reveal patterns in energy consumption, facilitating a deeper understanding of underlying trends and variances.

- **Seasonality:** Bar plots were employed to illustrate variations in energy consumption across different seasons. As shown in Fig. 7, energy consumption peaks during the spring months, followed closely by winter, likely due to increased heating demands associated with colder temperatures in winter and heightened activity in spring. This seasonal analysis provides critical insights for forecasting energy needs and optimizing resource allocation throughout the year.
- **Monthly Energy Usage:** Box plots displayed the distribution of energy consumption for each month, highlighting significant spikes in energy usage during colder months. Figure 8 reinforces the seasonal trends previously identified, showcasing the variability across months.
- **Energy Consumption Distribution:** The distribution of 'Actual_Trev' values was analyzed, revealing a peak at 15.62 MW, indicating a common consumption level. This analysis confirms the effectiveness of the preprocessing steps and provides insight into energy consumption patterns. Figure 9 illustrates the distribution and density of energy consumption, indicating that the highest density occurs at 15.62 MW, with values ranging from approximately 7.81 MW to 17.22 MW.

3.1.5 Train, validation, and test dataset

To effectively evaluate the LSTM model, the dataset was divided into three distinct subsets:

- **Training Set:** This subset was utilized to train the LSTM model on historical energy

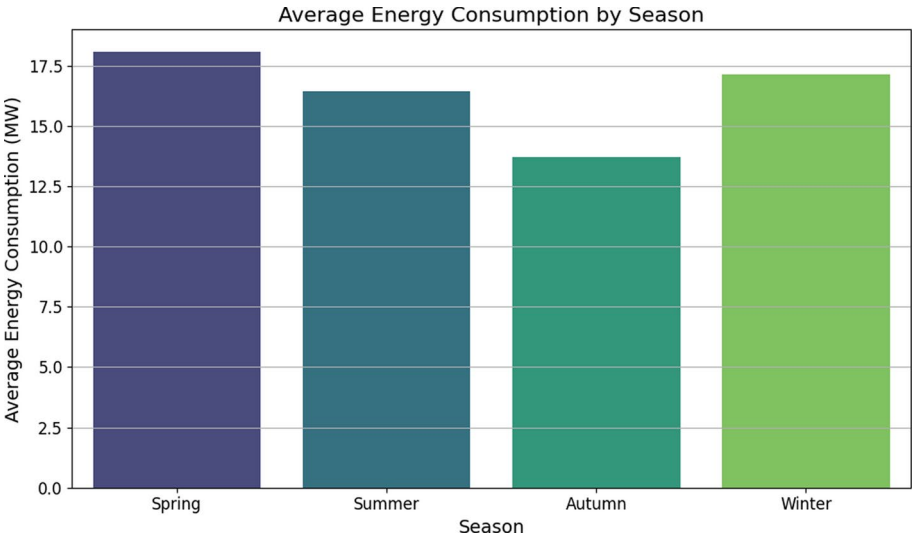

Fig. 7 Average energy consumption by season. The bar plot highlights the significant increase in energy usage during spring, followed by winter, indicating a strong correlation between seasonal changes and energy demand

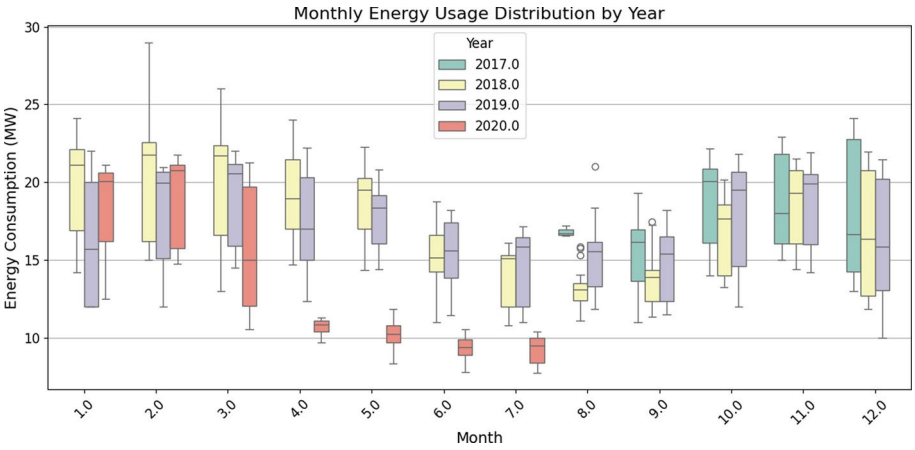

Fig. 8 Monthly energy usage distribution. The box plot illustrates the range of energy consumption for each month, highlighting significant spikes during colder months

consumption data, enabling the model to learn inherent patterns and relationships.

- **Validation Set:** The validation set served to fine-tune hyperparameters and monitor model performance during training. This step is critical for preventing overfitting, ensuring a balance between fitting the training data and generalizing to new, unseen data.
- **Test Set:** This subset was reserved for the final evaluation of the model's performance. Using a separate test set allows for an assessment of the model's generalization ability on data it has not encountered during training or validation.

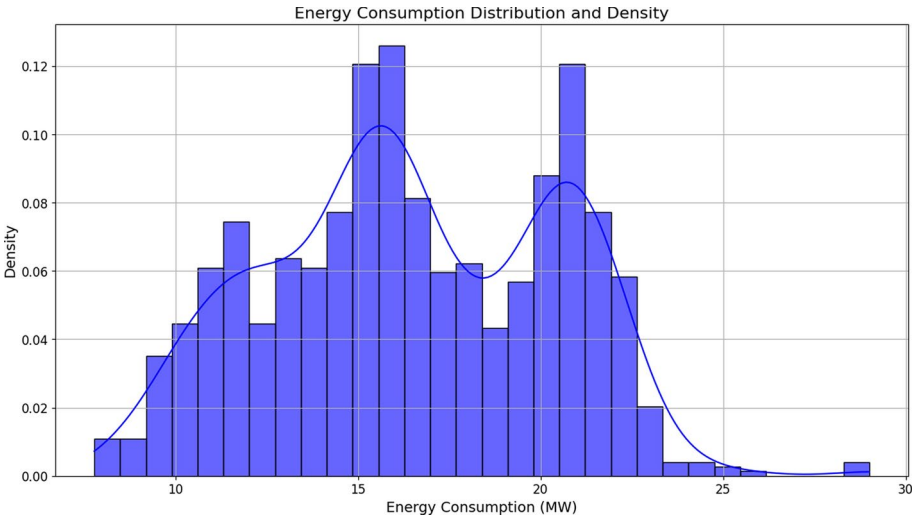

Fig. 9 Energy consumption distribution and density. This figure illustrates the distribution of energy consumption values for the 'Trevithick' building, with the highest density observed at 15.62 MW

These strategic splits were instrumental in confirming the model's capability to generalize effectively to future energy consumption data, rather than merely memorizing the training dataset.

3.1.6 LSTM mechanism

The underlying mechanism of the LSTM includes equations that govern the flow of information through its gates:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad \text{(Forget Gate)} \tag{1}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{(Input Gate)} \tag{2}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{(Output Gate)} \tag{3}$$

These gates work together to capture long-term dependencies in time series data, making the LSTM model particularly effective for prediction tasks.

3.1.7 Model structure

The architecture of the LSTM model comprised three primary components:

- **LSTM Layer:** The core of the model consisted of a single LSTM layer with 50 units. This layer was essential for capturing temporal patterns in the energy consumption data, enabling the model to learn from both short- and long-term dependencies.
- **Dropout Layer:** A dropout layer with a 20% dropout rate was integrated to mitigate the risk of overfitting. By randomly dropping a fraction of the units during training, the

model was encouraged to learn more robust features, thereby improving its generalization ability.

- **Dense Layer:** The final layer was a fully connected Dense layer that produced the output for energy consumption prediction. This layer enabled the model to synthesize learned information from previous layers into a single prediction.

This model structure effectively addressed the complexities of time series forecasting in energy consumption, balancing model complexity with generalization capabilities.

3.1.8 Model performance

The performance of the LSTM model in predicting daily energy consumption is illustrated through two key graphical representations. Figure 10 compares actual energy consumption (in blue) with the LSTM model's predictions (in orange) over time. This visual representation underscores the model's ability to closely follow actual consumption trends, demonstrating an accuracy of 95.87% on the validation set. Such proximity indicates effective learning and generalization, which are essential for reliable forecasting. Additionally, Fig. 11 depicts the training and validation loss over the epochs. The training loss shows a consistent decline, while the validation loss decreases with some fluctuations. This trend suggests that the model is effectively minimizing error without significant overfitting, as both losses stabilize towards the end of training.

These results underscore the robustness of the LSTM model's predictive capabilities and its effectiveness in capturing underlying patterns in energy consumption data. The high accuracy achieved indicates that the model can be reliably used for forecasting purposes in energy management applications. Moreover, the observed loss trends suggest that further tuning and optimization could enhance performance, especially in real-time applications. Continued evaluation and refinement of the model will be essential to maintain its accuracy and adaptability to changing energy consumption behaviors in diverse environments.

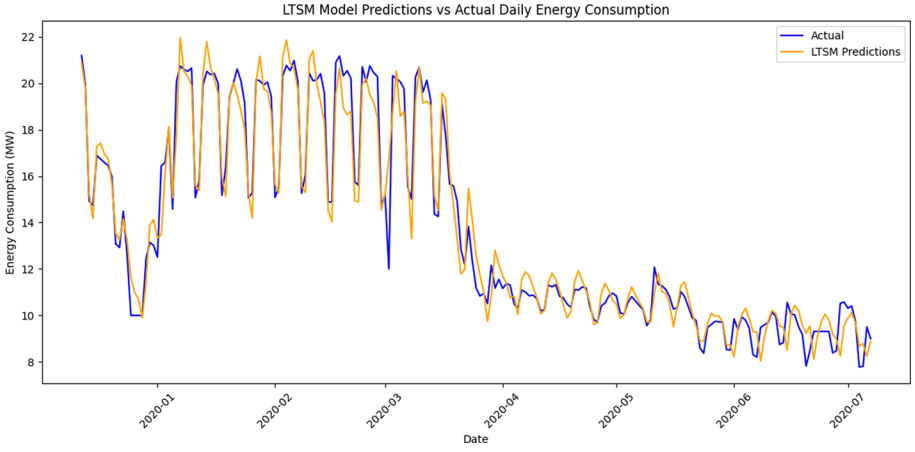

Fig. 10 LSTM baseline model predictions vs. actual energy consumption. This figure displays the baseline model's performance, with the blue line for actual values and the orange line for predictions over time

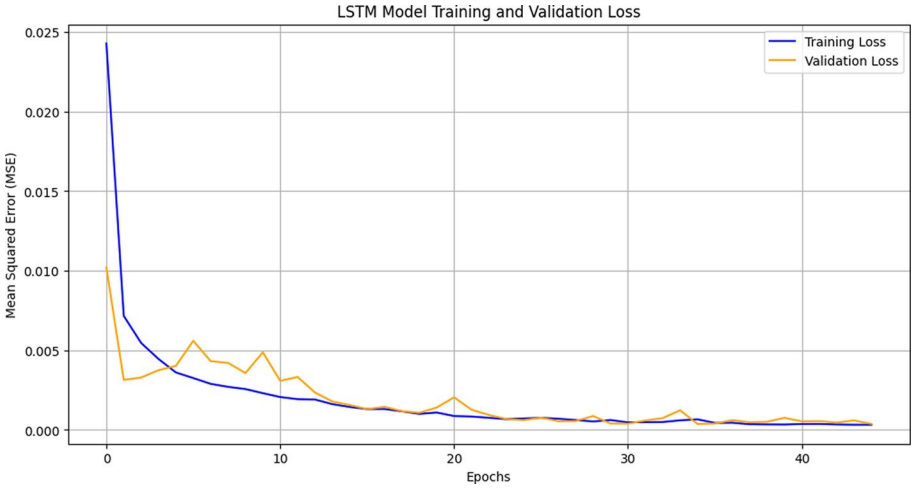

Fig. 11 LSTM model training and validation loss. This figure illustrates the training and validation loss over epochs, highlighting the model's learning progress and stability

3.1.9 Future predictions for energy consumption

To forecast energy consumption for the next 60 days using the trained LSTM model, we implemented the `create_future_data` function. This function generates predictions by following a systematic approach:

1. **Input Preparation:** The last data point from the scaled dataset is reshaped to conform to the model's input requirements, ensuring it has the shape $(1, \text{timesteps}, \text{features})$.
2. **Iterative Prediction:** A loop runs for the specified number of days (60 in this case). For each iteration:
   - The model predicts the next energy consumption value based on the current data.
   - The predicted value is appended to a list for future analysis.
   - The input data is updated by discarding the oldest timestep and adding the newly predicted value, maintaining the required input shape.
3. **Inverse Transformation:** Once all predictions are generated, the values are transformed back to their original scale using the inverse of the scaling applied during training.

The predicted energy consumption values show an initial increase, peaking around Day 10, followed by a decline toward Day 20. After this, values fluctuate with an upward trend by Day 60. This cyclical pattern in energy demand can significantly aid in resource planning and management decisions.

3.2 LSTM models optimization for energy consumption

This study evaluates three LSTM model variations: Standard LSTM, Pruned LSTM, and Quantized LSTM, focusing on their accuracy and inference speed. The Standard LSTM model achieved an accuracy of 92.43%, serving as a baseline. After pruning, accuracy increased slightly to 92.97%, attributed to the regularization effect that reduces overfitting. In contrast, the Quantized LSTM model exhibited a decrease in accuracy to 90.25%, due to the introduction of quantization noise from lowering weight precision. Nonetheless, this model remains viable where efficiency is prioritized. Figure 12 shows the model predictions compared to actual daily energy consumption, illustrating the performance differences among the models.

The model predictions reveal that the Standard LSTM closely aligns with actual consumption, demonstrating its high accuracy. The Pruned LSTM maintains a similar trend, indicating effective learning despite reduced complexity. Although the Quantized LSTM shows a decline in accuracy, it still captures essential consumption patterns, highlighting its efficiency in resource-constrained scenarios. In addition to accuracy, inference speed was analyzed to assess practicality for real-time applications. The Standard LSTM had an inference time of 0.994 s per sample, which is impractical for edge devices. After pruning, this time decreased to 0.566 s, a 43% reduction due to reduced model complexity. As shown in Fig. 13, quantization provided the most significant improvement, achieving an inference time of only 0.095 s per sample, reflecting a 90% speedup by utilizing lower-precision arithmetic.

This comparison highlights the substantial speed improvements achieved through pruning and quantization. The Quantized LSTM's remarkable reduction in inference time demonstrates its suitability for real-time applications, particularly in low-power environments, making it a strong candidate for energy management systems. Weight distribution analysis further elucidates the models' characteristics. The Standard LSTM exhibited a wide spread of weight values, indicating diverse learned relationships, whereas pruning resulted in many weights being reduced to zero, concentrating remaining weights around small magnitudes.

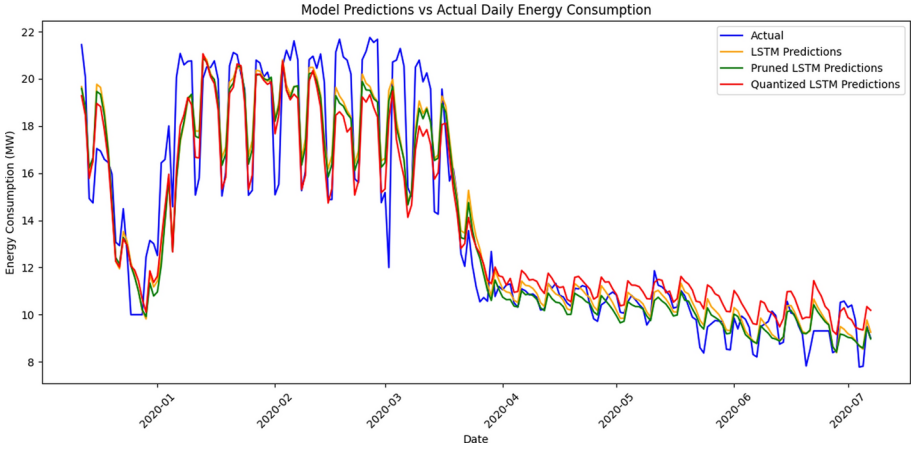

Fig. 12 Models predictions vs. actual daily energy consumption. This figure compares predictions from three LSTM variations: Standard, Pruned, and Quantized LSTMs

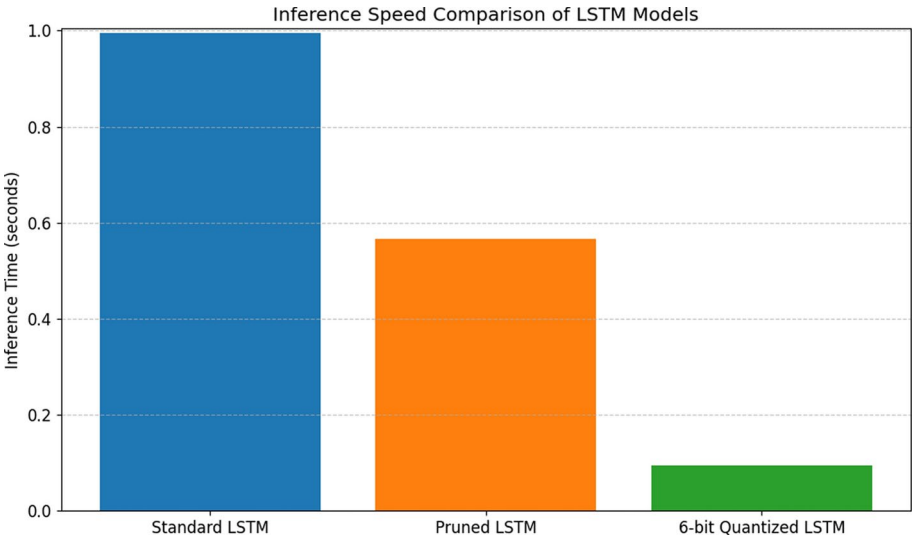

Fig. 13 This figure compares the inference speeds of Standard, Pruned, and Quantized LSTMs. Quantization delivers the best performance at 0.095 s per sample, allowing efficient deployment on edge devices

This sparsity, illustrated in Fig. 14, allows the pruned model to maintain accuracy through effective regularization. In contrast, the Quantized LSTM demonstrated a more uniform weight distribution due to reduced bit precision, which, while improving efficiency, contributed to the observed accuracy drop.

The weight distribution analysis shows that the Pruned LSTM effectively eliminates many weights, allowing the model to concentrate on significant features, thus enhancing generalization. Conversely, the Quantized LSTM's uniform distribution reflects the trade-offs involved in reducing precision for efficiency. When comparing the models, each optimization technique has distinct advantages. The Standard LSTM offers high accuracy but is less viable for real-time applications due to its computational cost. Pruning strikes a balance by slightly improving accuracy while significantly reducing inference time, making it suitable for environments with limited resources. Conversely, the Quantized LSTM excels in inference speed, suitable for deployment in low-power scenarios, despite a modest accuracy reduction.

3.3 NAS and edge deployment of optimized LSTM models

The NAS LSTM model significantly outperforms the Standard LSTM in prediction accuracy, achieving a Mean Squared Error (MSE) of 0.00018 compared to 0.00305, and an R^2 value of 0.9960 versus 0.9589 (Table 4). This improvement stems from the NAS process's rigorous search for optimal hyperparameters and architectures, though it increases training time (139.49 s vs. 4.83 s). Despite this trade-off, the NAS LSTM exhibits slightly faster inference (0.0676 s vs. 0.0714 s) and is marginally larger (0.31 MB vs. 0.28 MB), which may impact deployment on resource-constrained devices. To deploy the NAS LSTM on FPGAs, quantization is required, potentially reducing accuracy-a key consideration when balancing performance and hardware constraints.

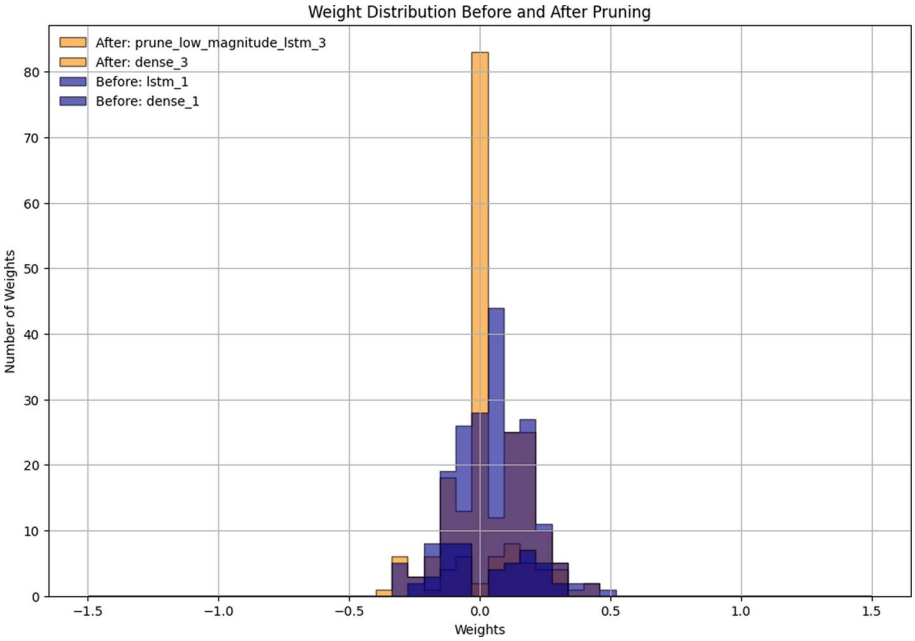

Fig. 14 Weight distribution before and after pruning. The Standard LSTM shows a wide weight spread, while pruning concentrates weights around small magnitudes for sparsity

Table 4 Comparison of standard LSTM and NAS LSTM performance metrics

Metric	Standard LSTM	NAS LSTM
Mean squared error (MSE)	0.00305	0.00018
R-squared (R^2)	0.9589	0.9960
Training time (s)	4.83	139.49
Inference time (s)	0.0714	0.0676
Model size (MB)	0.28	0.31

Cloud-based solutions, while effective for large-scale data management, introduce challenges such as latency, cybersecurity risks, and high data transfer costs (Yanamala 2024). To address these limitations, we evaluated optimized LSTM models on edge hardware platforms, including FPGAs and GPUs. Figure 15 shows that the Pynq Z1 FPGA achieves a superior inference speed of 0.002574 s, compared to the Intel(R) Xeon(R) W7-3445 CPU at 0.289659 s and the NVIDIA GeForce RTX 4070 Ti GPU at 0.127011 s, making it ideal for real-time applications like adaptive HVAC control and fault detection.

Quantization further enhances edge compatibility by reducing the Baseline Model's inference size from 70.55 to 27.69 KB and weight size from 13.50 to 1.19 KB (Fig. 16)

The proposed FastML model has generated a low inference time (0.002574 s) which makes it ideal for edge deployment, where minimizing latency is critical, such as in dynamically adjusting HVAC systems based on occupancy patterns. The model size of the proposed quantized model (27.69 KB inference size, 1.19 KB weight size) further supports edge deployment feasibility. Overall, FastML has advanced compatibility with edge or fog environments allowing for localized data processing closer to data capture or actuation points.

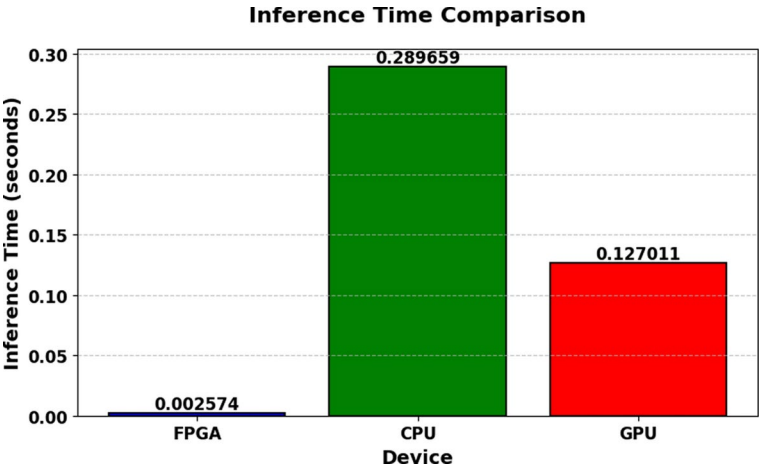

Fig. 15 Inference speed comparison of optimized LSTM models across hardware platforms. FPGA acceleration provides a significant advantage in real-time applications

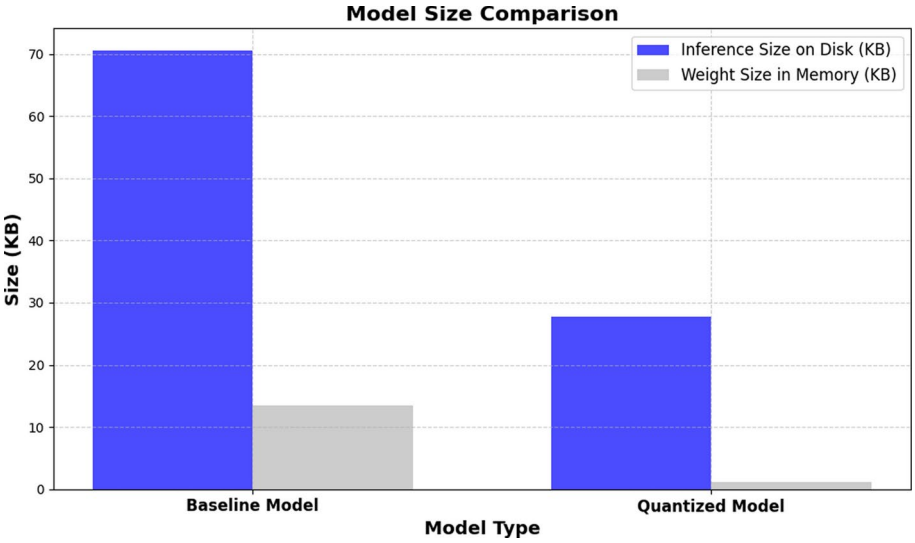

Fig. 16 Comparison of model sizes between Baseline and Quantized Models, focusing on inference disk storage and weight size

This deployment enhances data locality, as analytics and decision-making occur on-site, reducing latency with the ability to provide near real-time signals to changing conditions in BMSs. By minimizing data transmission to the cloud, FastML alleviates bandwidth constraints, making efficient use of network resources and reducing operational costs. Furthermore, processing data at the edge mitigates cybersecurity risks, as sensitive information can be analyzed locally without being transmitted over networks. This in-site analysis enhances data privacy and security, ensuring compliance with regulations while maintaining the integrity of occupant data.

3.4 Transformation to HLS4ML

In this experiment, we implemented and compared a machine learning model using two approaches: a traditional Keras implementation and a hardware-optimized version using HLS4ML. The goal was to evaluate the performance of the HLS4ML model, which was derived from the predictions generated by the LSTM model, in terms of accuracy and resource efficiency compared to the baseline Keras model. The Keras model achieved an accuracy of 95.01%, while the HLS4ML model attained 92.35%, reflecting a slight drop of 2.66%. This decrease in accuracy is attributed to the fixed-point quantization employed by HLS4ML, as opposed to the floating-point precision used in Keras. Fixed-point arithmetic introduces quantization error, leading to a marginal reduction in accuracy but allowing for significantly more efficient computations on hardware platforms. The model architecture used for this comparison consists of three dense layers. The input layer receives a 20-dimensional input vector, followed by fully connected dense layers that progressively reduce the dimensionality to a single output representing the model's prediction. Both the Keras and HLS4ML implementations employed this architecture to ensure a fair comparison. The HLS4ML model was configured with specific parameters: the precision was set to $\mathrm{ap\_fixed}<16,6>$, with 16 bits total and 6 bits allocated for the integer part. The reuse factor, determining the level of resource reuse in the hardware, was set to 1, minimizing latency by performing computations in parallel at the cost of increased resource usage. The optimization strategy focused on reducing latency during inference, while the BRAM factor, specifying the available block RAM resources, was set to 1,000,000,000. Trace output was disabled to minimize runtime overhead. Each layer's weights, biases, and outputs were quantized using the same fixed-point precision, ensuring consistent numerical representation throughout the model.The results of the predictions made by both models are shown in Fig. 17. The graph indicates that while the Keras model consistently performs at a higher accuracy, the HLS4ML model still delivers competitive results. This performance trade-off highlights the effectiveness of using hardware optimization for deployment in environments where computational resources are limited.

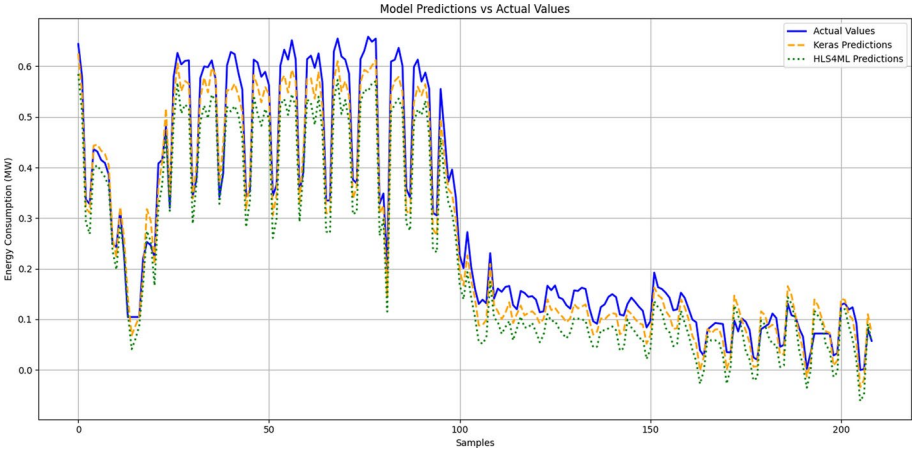

Fig. 17 Comparison of HLS4ML and Keras model predictions. This figure shows the accuracy of both models, with HLS4ML reflecting a 2.66% drop compared to the Keras model

The slight reduction in accuracy can be considered an acceptable compromise for the gains in resource efficiency and computational speed offered by the HLS4ML model. These findings suggest that for applications where latency and resource utilization are critical, the HLS4ML framework presents a viable solution despite the marginal loss in predictive performance. Future work could explore advanced quantization techniques or model architectures to further bridge the accuracy gap while maintaining the benefits of hardware acceleration.

3.4.1 HLS4ML model configuration

The HLS4ML model was configured with a 16-bit fixed-point precision ($\mathrm{ap\_fixed} < 16,6 >$), allocating 6 bits for the integer part. The reuse factor was set to 1, thereby minimizing latency by allowing parallel computations, although this increased resource usage. The optimization of the model focused on reducing latency during inference, with a Block RAM (BRAM) factor of 1,000,000,000. Additionally, trace output was disabled to decrease runtime overhead. All layers utilized the same fixed-point precision to ensure consistent numerical representation.

The structure of the HLS4ML model, as illustrated in Fig. 18, showcases the overall architecture and highlights the parallel processing capabilities that contribute to its performance efficiency. This configuration is crucial for achieving optimal inference times, especially in applications requiring rapid decision-making.

3.4.2 Weight profiling

Weight profiling was conducted both before and after optimization to analyze the distribution and sparsity of the model's weights. Initially, the weight distribution was broad, with minimal sparsity, as depicted in Fig. 19.

After the HLS4ML optimization, the weight distribution became more concentrated, with non-essential weights pruned to zero, thereby enhancing sparsity. This increased sparsity not only reduced model complexity but also improved hardware efficiency, as shown in Fig. 20.

The comparison of weight distributions before and after optimization underscores the effectiveness of the pruning technique, highlighting significant improvements in both model performance and resource utilization.

3.4.3 Model architecture visualization

The model topology was visualized to illustrate the data flow through the network, showcasing how each layer transforms the input features into the final prediction. Below is a summary of the key layers in the model (Table 5):

This visualization emphasizes the flow of information and the consistent use of fixed-point precision throughout the model, ensuring that the architecture is both efficient and effective for deployment.

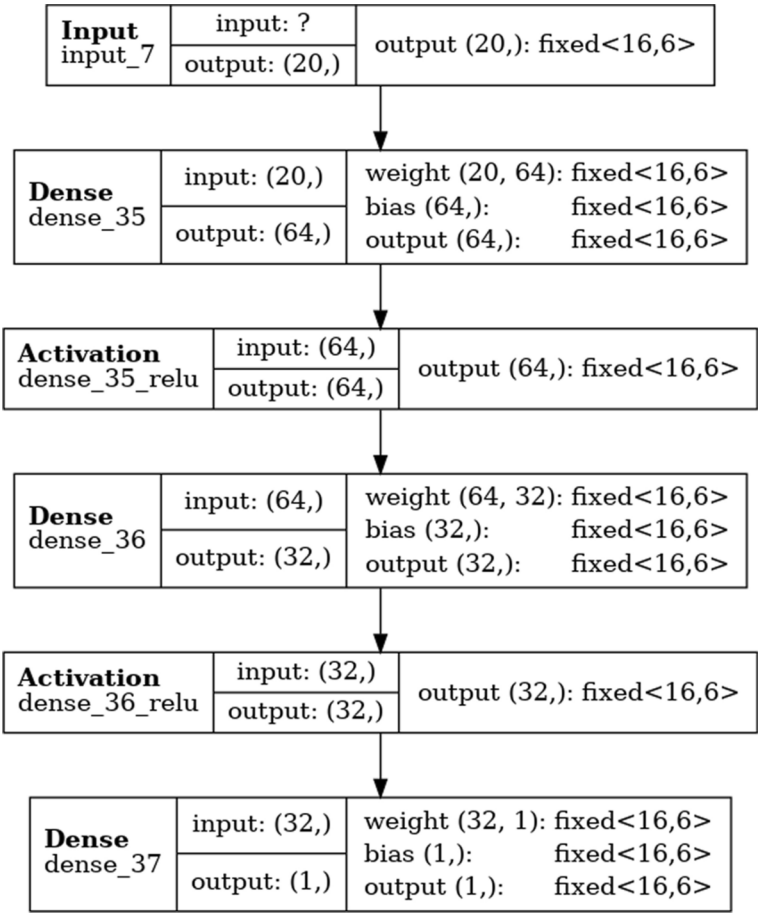

Fig. 18 HLS4ML model summary. This figure details the model's architecture, including the input layer, hidden layers, and output layer for energy prediction. It also indicates the precision used throughout the model

4 Discussion

This study explores enhancing BMSs using FastML techniques. Our findings indicate that FastML can significantly improve BMS performance, automation, and efficiency through pruned and quantized models. The Pruned LSTM model achieved a 43% increase in inference speed with an accuracy of 92.97%, making it suitable for resource-constrained environments like IoT devices. The Quantized LSTM resulted in a 90% reduction in inference time, crucial for real-time energy management, with only a minor compromise in accuracy. We primarily utilized LSTM models for energy consumption forecasting, evaluating three optimized variations: Standard, Pruned, and Quantized. These models effectively capture temporal dependencies in energy data, with the pruned and quantized versions tailored for resource-efficient applications. The integration of the HLS4ML framework enabled hardware acceleration on platforms like FPGAs, significantly boosting inference speed for

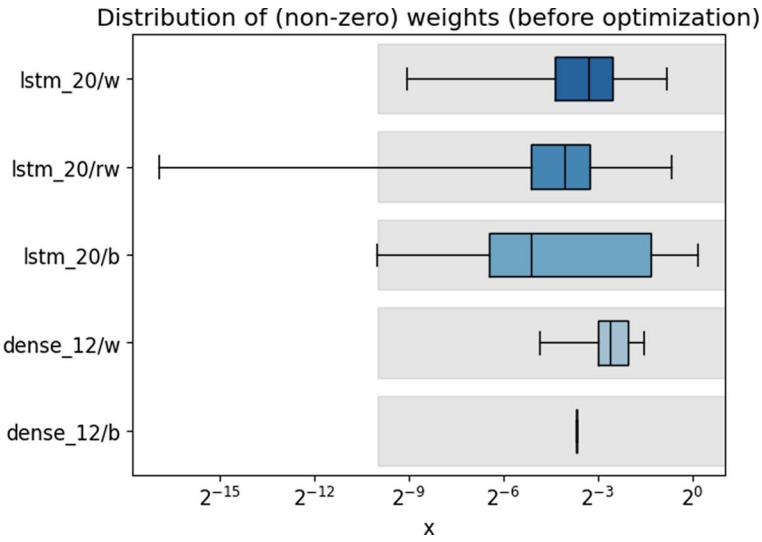

Fig. 19 Weight profiling before optimization

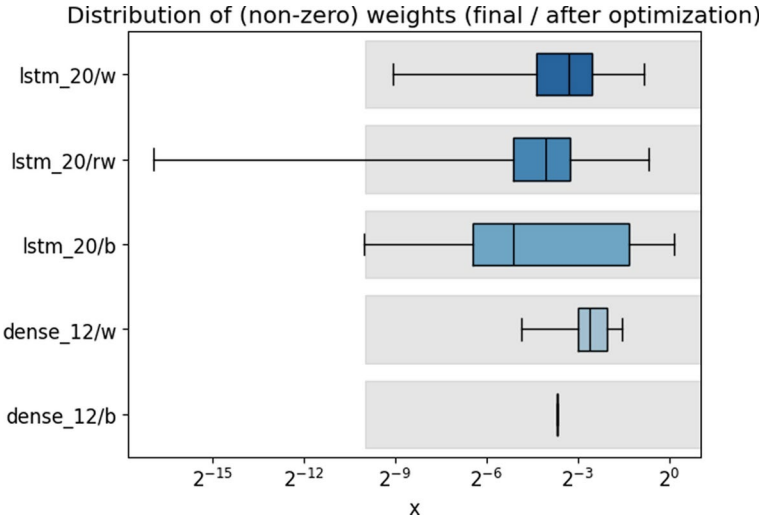

Fig. 20 Weight profiling after optimization

energy forecasting tasks. Our research involved the creation, deployment, and testing of FastML models in a real-world case study at the Queen's Building. This process included optimizing LSTM models and transforming them into hardware-accelerated versions via HLS4ML, improving inference speed while maintaining acceptable accuracy. Although the fixed-point quantization method resulted in a slight accuracy drop (92.35%), it significantly enhanced computational efficiency, making the models suitable for real-time energy management. Testing confirmed their effectiveness in handling real-world energy consumption data. While the focus of this study has been on technical advancements, broader factors such

Table 5 Overview of key model layers and precision settings

Layer name	Layer type	Precision	Trace	Input shape
input_7	InputLayer	ap_fixed <16,6>	False	[(None, 20)]
dense_35	Dense	ap_fixed <16,6>	False	[(None, 20)]
dense_35_relu	ReLU	ap_fixed <16,6>	False	[(None, 64)]
dense_36	Dense	ap_fixed <16,6>	False	[(None, 64)]
dense_36_relu	ReLU	ap_fixed <16,6>	False	[(None, 32)]
dense_37	Dense	ap_fixed <16,6>	False	[(None, 32)]
dense_37_linear	Linear	ap_fixed <16,6>	False	[(None, 1)]

This table includes each layer's name, type, precision format, tracing status, and input shape

as regulatory constraints, building codes, and evolving energy standards are critical for real-world adoption. For instance, compliance with mandates like LEED for energy efficiency or GDPR for data protection could influence the feasibility and scalability of these solutions, while regulatory incentives may enhance their adoption. FastML enables a rapid model development and deployment while ensuring compliance with regulatory constraints and evolving energy standards. FastML facilitates swift machine learning model training and optimization, allowing for quick adaptation to new building codes and energy regulations. Meanwhile, HLS4ML converts high-level machine learning models into hardware-efficient implementations, ensuring real-time performance and reliability, crucial for meeting stringent safety and performance standards. To address non-stationary environments, our model incorporates strategies to enhance model adaptability. Temporal features like season, month, and lagged energy consumption capture gradual changes in climate and usage patterns. Additionally, the model can be periodically retrained with updated data to adapt to evolving load profiles or energy policies. FastML models can adapt to dynamic building conditions such as load profiles, occupant requirements or external policies by leveraging real-time data assimilation and adaptive learning algorithms. The HLS4ML facilitates rapid processing at the edge, enabling immediate adjustments with continuous learning capabilities to support inference based on the incoming data, allowing it to recalibrate control strategies when user parameters such as occupancy or comfort change. Furthermore, many machine learning models, including FastML, rely on high-quality datasets. The lack of sufficiently large and diverse datasets for training and validation is a concern, particularly when energy consumption data is not reported. This data gap can hinder model accuracy and reliability, emphasizing the need for improved data collection and preprocessing techniques. Looking ahead, enhancing the real-time adaptability of FastML models to changing building conditions, such as occupancy levels and weather changes, will be crucial for optimizing energy management. Seamless integration of FastML solutions with existing BMS infrastructure is also essential, highlighting the need for standardized protocols to facilitate integration without complete system overhauls. Prioritizing user-friendly interfaces and decision support tools for building operators and occupants will enhance engagement and effectiveness in energy management. Robustness against anomalies in data, such as sensor malfunctions or unexpected environmental changes, is critical for model reliability. Further research could

explore anomaly detection mechanisms to strengthen this aspect. As BMS increasingly integrate machine learning, evaluating the environmental impact of deploying these technologies-particularly regarding energy consumption and resource usage-should also be a focus.

In summary, refining these models through techniques such as layer fusion and advanced pruning strategies will be vital for enhancing scalability and integration with smart grid technologies. Addressing these challenges is crucial for maximizing the practical impact of FastML solutions in improving energy efficiency and operational performance in BMS.

5 Final remarks and future roadmap

This research showcases the application of HLS4ML-based FastML techniques in BMSs through a case study at the Queen's Building, utilizing FPGA hardware enabled by HLS4ML. It simplifies the deployment of machine learning models, making them more accessible to operators and reducing setup time. Optimization methods like pruning and quantization enhance real-time energy management on resource-constrained devices, while FastML excels in speed and efficiency, making it suitable for rapid scenarios like fault detection. Furthermore, FastML effectively handles predictive uncertainty, model drift, and unexpected conditions based on a robust adaptive learning capability and hardware-efficient implementations. The advantage of HLS4ML facilitates quick inference at the edge, allowing for immediate adjustments to control strategies in response to unforeseen events like lockdowns or severe weather. The system can incorporate feedback loops that monitor performance metrics, enabling it to detect model drift and recalibrate accordingly. The dataset utilized for this study is sourced from Queen's Building, an educational facility located in South Wales. To address challenges such as data gaps, sensor faults, and irregular maintenance logs, FastML provides key mechanisms to address data gaps and privacy by employing advanced data imputation techniques to handle sensor faults and gaps in data collection, ensuring that analyses remain robust and reliable even in the presence of incomplete information. Moreover, FastML can detect and compensate for irregular maintenance logs, enhancing the accuracy of predictive models and operational insights. Regarding privacy, FastML integrates privacy-preserving methods, such as data anonymization and encryption, to protect sensitive information while still enabling effective data analysis by using federated learning approaches, allowing models to be trained on decentralized data sources without compromising individual privacy. FastML is specifically designed to accelerate and optimize processes for speed and efficiency, making it particularly well-suited for scenarios requiring rapid response and decision-making, such as fault detection. In this study, FastML was effectively applied to energy forecasting in BMSs, demonstrating its real-time capabilities, effective speed, resource efficiency, and strong predictive performance. The hardware integration with hls4ml can enable efficient inference at the edge to capture timely alerts and data-driven adjustments for critical assets. Beyond fault detection, FastML can be integrated with multi-objective optimization techniques, particularly GAs, to balance competing objectives within BMS applications. Such integration can minimize energy consumption while maintaining occupant comfort and operational efficiency. GAs provide a robust framework for identifying Pareto-optimal solutions, enabling informed decision-making that reconciles these competing demands. FastML has the ability to effectively address the trade-offs between energy savings, occupant comfort, and real-time responsiveness in build-

ing operations through advanced ML and hardware optimization techniques. FastML utilizes data-driven insights and predictive analytics to analyze occupancy patterns and energy usage, enabling proactive adjustments in HVAC systems that enhance comfort while minimizing energy consumption. Simultaneously, HLS4ML translates these machines learning models into efficient hardware implementations, ensuring rapid decision-making and real-time responsiveness. By enabling multi-objective optimization and incorporating occupant preferences, these tools create a dynamic feedback loop that balances the need for energy efficiency with the imperative of occupant comfort, ultimately leading to smarter and more sustainable building management practices. Furthermore, FastML provides significant advantages for user-centric solution in buildings by integrating with multi-objective optimization techniques such as GAs and MPC. This integration enables the formulation of adaptive control policies that effectively balance multiple objectives, including energy efficiency, occupant comfort, and operational costs. By leveraging real-time data on human behaviour and preferences, FastML can inform these optimization processes, ensuring that control strategies are responsive to occupant needs with feedback integration for continuous refinement of policies based on user input, enhancing overall satisfaction. Policy constraints, such as regulatory requirements and sustainability goals, can be also seamlessly integrated as constraints into the optimization framework, ensuring that BMS operations remain compliant while maximizing performance. By integrating with advanced algorithms such as GAs, neural networks and reinforcement learning, FastML can capture the complex interdependencies between environmental controls and occupant actions. For instance, it can analyze historical data to identify patterns in how occupancy influences HVAC load and lighting requirements, while simultaneously considering security protocols that might alter access and energy usage. Incorporating real-time data from sensors and IoT devices enables the system to dynamically adjust to changing conditions, such as varying occupancy levels or external weather factors. FastML can learn from occupant behavior and preferences, optimizing energy efficiency and comfort while maintaining security. Scalability is another key strength of FastML, allowing deployment at district or city-wide levels. The framework enables efficient and distributed model deployment alongside data aggregation capabilities. HLS4ML facilitates the conversion of ML models into hardware-efficient implementations, enabling real-time processing at the network edge for each building. This localized deployment minimizes latency and bandwidth requirements while allowing buildings to operate autonomously based on real-time data. Aggregating data from these distributed models into a central system allows for comprehensive analytics, facilitating city-wide insights and trend identification. Coordinated demand response strategies can leverage aggregated data to balance loads across buildings, further enhancing overall energy efficiency. FastML also supports communication between remote building systems for synchronized interventions, aligning energy use with city-wide sustainability goals and fostering a more resilient urban environment. Additionally, FastML leverages techniques such as feature fusion and deep learning architectures, which can synthesize information from various data sources to enhance predictive accuracy and decision-making capabilities. HLS4ML optimizes these models for hardware efficiency, enabling real-time processing at the edge. This combination allows BMS to dynamically adapt to varying conditions, providing a holistic view of building performance and facilitating informed responses to maintain occupant comfort and energy efficiency. Future work should develop advanced methodologies for time-series alignment, robust imputation techniques for missing data, and blockchain-based frame-

works to ensure data integrity and traceability. Enhancing data reliability in this manner will strengthen the robustness of FastML models, improving their effectiveness in real-world deployments. Finally, the lifecycle costs of adapting FastML for BMSs include expenses for model development, deployment, maintenance, retraining, and hardware upgrades. Maintenance costs involve regular calibration and upkeep of sensors and monitoring devices. The ongoing power consumption of FPGA hardware is an important factor, although it is likely to be lower than that of traditional systems due to the energy efficiency of FPGAs. Periodic updates to machine learning models and HLS4ML software will incur costs. Retraining and upgrade expenses will be incurred when building conditions or requirements change and resources will be needed for retraining models. There may also be hardware upgrades necessary to support more complex models or to handle increased data volumes as the system expands to cover more areas or functions. The HLS4ML framework provides configurable parameters that allow users to adjust the balance between latency, throughput, power consumption, and resource usage, hence, reducing costs while ensuring optimal performance for their specific applications.

6 Conclusion

The application of ML within BMSs has significantly improved energy efficiency and occupant comfort. However, conventional ML techniques often fall short in meeting the stringent timing and resource constraints required in BMS applications. FastML emerges as a pivotal approach, enhancing the performance of ML models in resource-constrained environments by accelerating inference and optimizing resource usage. This study presents a comprehensive review of ML and AI applications for BMSs, complemented by a case study using the LSTM model for energy forecasting in an educational building. By employing FastML techniques, the standard LSTM model was adapted for enhanced generalization, accuracy, and inference speed-qualities essential for real-time, performance-sensitive applications. Specifically, the pruned LSTM model achieved an accuracy of 92.97%, indicating effective regularization that mitigates overfitting. In comparison, the quantized LSTM, with a slightly lower accuracy of 90.25%, demonstrated notable improvements in inference speed, making it ideal for deployment in real-time, resource-limited environments. This trade-off analysis between accuracy and speed provides valuable insights for practical FastML deployment. Additionally, this paper proposes the use of HLS4ML, a high-level synthesis framework, as an efficient solution for implementing ML models on hardware platforms. While the HLS4ML model achieved comparable accuracy (92.97%) to its Keras counterpart, it offered substantial gains in hardware efficiency, particularly when deployed on FPGA and ASIC platforms. By utilizing fixed-point arithmetic, HLS4ML enables low-latency, high-throughput inference, which is well-suited for real-time BMS applications. A key consideration is the lifecycle management of FastML models. Specifically, FastML can be validated, maintained, and updated long-term through hardware-accelerated solutions and pruned or quantized models by implementing a robust framework that emphasizes interoperability and adaptability. Continuous validation is facilitated through real-time performance monitoring and feedback loops that assess model accuracy against established benchmarks, enabling prompt detection of performance drifts. Moreover, the retraining process can be automated and scheduled, ensuring model relevance with minimal manual intervention. Additionally,

to facilitate integration with legacy systems and proprietary BMSs, FastML and HLS4ML employ standardized APIs and middleware solutions, allowing seamless data exchange and control commands. Furthermore, FastML ensures continuous refinement and adaptation, mitigating model obsolescence and preserving long-term value through several mechanisms. For instance, it supports incremental learning, enabling models to update with new data without complete retraining, thus maintaining relevance as building conditions evolve. Similarly, the modular architecture of HLS4ML allows for easy integration of new hardware and algorithm updates, accommodating advancements in technology and utility policies. Continuous performance monitoring further enables real-time assessment and proactive adjustments based on feedback and changing conditions within the building ecosystem. Finally, dynamic retraining capabilities ensure models are regularly updated with new data, while configurable parameters in HLS4ML permit customization for specific requirements, such as energy consumption patterns. Looking ahead, future research may explore advanced optimization techniques, including layer fusion, enhanced pruning strategies, and refined quantization methods, to minimize accuracy loss while maximizing hardware efficiency. With these advancements, HLS4ML has the potential to become a cornerstone framework for deploying machine learning models in resource-constrained BMS environments, significantly advancing the field of energy management and control in real-world applications

Author Contributions M.M. conducted the case study and provided the experimental results. I.P. contributed to the overall conceptualization, methodology, and editing of the manuscript. Both authors reviewed and approved the final version of the manuscript.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Aarrestad T, Loncar V, Ghielmetti N, Pierini M, Summers S, Ngadiuba J, Petersson C, Linander H, Iiyama Y, Di Guglielmo G et al (2021) Fast convolutional neural networks on FPGAs with hls4ml. Mach Learn Sci Technol 2(4):045015

Abdelaziz A, Santos V, Dias MS (2023) A proposed intelligent model with optimization algorithm for clustering energy consumption in public buildings. Int J Adv Comput Sci Appl 10:100

Abdelaziz A, Santos V, Dias MS, Mahmoud AN (2024) A hybrid model of self-organizing map and deep learning with genetic algorithm for managing energy consumption in public buildings. J Clean Prod 434:140040

Abdullah MA, Razali MA, Abd Rahman R, Ishak M, Md Som M, Yusop A, Mohd Amin M, Abdul Hadi N (2022) Review of automation and energy monitoring system for air-conditioning and mechanical ventilation (ACMV) in building Malaysia. J Adv Mech Eng Appl. https://doi.org/10.30880/jamea.2022.03.01.002

Abuimara T, Hobson BW, Gunay B, O'Brien W, Kane M (2021) Current state and future challenges in building management: practitioner interviews and a literature review. J Build Eng 41:102803. https://doi.org/10.1016/j.jobe.2021.102803

Agouzoul A, Simeu E, Tabaa M (2022) Building energy consumption enhancement using a neural network based model predictive control synthesis in FPGA. In: 2022 International conference on microelectronics (ICM), December. IEEE, pp 262–265

Ahn KU, Park CS (2020) Application of deep q-networks for model-free optimal control balancing between different HVAC systems. Sci Technol Built Environ 26(1):61–74

Akbar MK, Amayri M, Bouguila N (2024) A novel non-intrusive load monitoring technique using semi-supervised deep learning framework for smart grid. Build Simul. https://doi.org/10.1007/s12273-023-1074-5

Alfaverh F, Denai M, Sun Y (2020) Demand response strategy based on reinforcement learning and fuzzy reasoning for home energy management. IEEE Access 8:39310–39321

Alqahtani T, Badreldin HA, Alrashed M, Alshaya AI, Alghamdi SS, bin Saleh K, Alowais SA, Alshaya OA, Rahman I, Al Yami MS (2023) The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. Res Soc Admin Pharm 19(8):1236–1242

Arias-Requejo D, Pulido B, Keane MM, Alonso-González CJ (2023) Clustering and deep-learning for energy consumption forecast in smart buildings. IEEE Access 11:128061–128080

Baird Z, Gunasekara I, Bolic M, Rajan S (2017) Principal component analysis-based occupancy detection with ultra wideband radar. In: 2017 IEEE 60th International midwest symposium on circuits and systems (MWSCAS)

Barbaresi A, Ceccarelli M, Menichetti G, Torreggiani D, Tassinari P, Bovo M (2022) Application of machine learning models for fast and accurate predictions of building energy need. Energies 15(4):1266

Bhagwat A, Dutta S, Jadoun VK, Veerendra AS, Sahu SK (2024) A customised artificial neural network for power distribution system fault detection. IET Gen Transmiss Distrib 18(11):2105–2118

Brandi S, Piscitelli MS, Martellacci M, Capozzoli A (2020) Deep reinforcement learning to optimise indoor temperature control and heating energy consumption in buildings. Energy Build 224:110225

Chen H, Luo H, Huang B, Jiang B, Kaynak O (2023a) Transfer learning-motivated intelligent fault diagnosis designs: a survey, insights, and perspectives. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.36227/techrxiv.21301533.v1

Chen Z, Xiao F, Guo F (2023b) Similarity learning-based fault detection and diagnosis in building hvac systems with limited labeled data. Renew Sustain Energy Rev 185:113612

Chen Z, Xiao F, Guo F, Yan J (2023c) Interpretable machine learning for building energy management: a state-of-the-art review. Adv Appl Energy 9:100123

Coraci D, Brandi S, Hong T, Capozzoli A (2023) Online transfer learning strategy for enhancing the scalability and deployment of deep reinforcement learning control in smart buildings. Appl Energy 333:120598

Cordeiro-Costas M, Labandeira-Pérez H, Villanueva D, Pérez-Orozco R (2024) NSGA-II based short-term building energy management using optimal LSTM-MLP forecasts. Int J Electr Power Energy Syst 159:110070

Deiana AM, Tran N, Agar J, Blott M, Di Guglielmo G, Duarte J, Harris P, Hauck S, Liu M, Neubauer MS et al (2022) Applications and techniques for fast machine learning in science. Front Big Data 5:787421

Deiana AM, Tran N, Agar J, Blott M, Di Guglielmo G, Duarte J, Warburton TK (2022) Applications and techniques for fast machine learning in science. Front Big Data 5:787421

Dey M, Rana SP, Dudley S (2018) Semi-supervised learning techniques for automated fault detection and diagnosis of HVAC systems. In: 2018 IEEE 30th international conference on tools with artificial intelligence (ICTAI)

Dey M, Rana SP, Dudley S (2020) A case study based approach for remote fault detection using multi-level machine learning in a smart building. Smart Cities 3(2):401–419

Digitemie WN, Ekemezie IO (2024) A comprehensive review of building energy management systems (BEMS) for improved efficiency. World J Adv Res Rev 21(3):829–841

Ding Z-K, Fu Q-M, Chen J-P, Wu H-J, Lu Y, Hu F-Y (2022) Energy-efficient control of thermal comfort in multi-zone residential HVAC via reinforcement learning. Connect Sci 34(1):2364–2394

Duarte J, Tran N, Hawks B, Herwig C, Muhizi J, Prakash S, Janapa Reddi V (2022a) FASTML science benchmarks: accelerating scientific edge ML [Poster]. In: Conference on machine learning and systems (MLSys 2022), Santa Clara, CA, USA, Aug–Sep 2022. https://doi.org/10.2172/1883578

Duarte J, Tran N, Hawks B, Herwig C, Muhizi J, Prakash S, Reddi VJ (2022b) FASTML science benchmarks: accelerating real-time scientific edge machine learning. arXiv preprint. arXiv:2207.07958

Durand D,. Aguilar J, R-Moreno MD (2022) An analysis of the energy consumption forecasting problem in smart buildings using lstm. Sustainability 14(20):13358
El-Maraghy M, Metawie M, Safaan M, Eldin AS, Hamdy A, El Sharkawy M, Abdelaty A, Azab S, Marzouk M (2024) Predicting energy consumption of mosque buildings during the operation stage using deep learning approach. Energy Build 303:113829
Etezadifar M, Karimi H, Mahseredjian J (2023) Non-intrusive load monitoring: comparative analysis of transient state clustering methods. Electr Power Syst Res 223:109644
Fährmann D, Jorek N, Damer N, Kirchbuchner F, Kuijper A (2022) Double deep q-learning with prioritized experience replay for anomaly detection in smart environments. IEEE Access 10:60836–60848
Fan C, Wu Q, Zhao Y, Mo L (2024) Integrating active learning and semi-supervised learning for improved data-driven HVAC fault diagnosis performance. Appl Energy 356:122356
Fan H, Ferianc M, Que Z, Liu S, Niu X, Rodrigues MR, Luk W (2022) FPGA-based acceleration for Bayesian convolutional neural networks. IEEE Trans Comput Aided Des Integr Circuits Syst 41(12):5343–5356
Fang P, Wang M, Li J, Zhao Q, Zheng X, Gao H (2023) A distributed intelligent lighting control system based on deep reinforcement learning. Appl Sci 13(16):9057
Farhadi B, You J, Zheng D, Liu L, Wu S, Li J, Li Z, Wang K, Liu S (2023) Machine learning for fast development of advanced energy materials. Next Mater 1(3):100025
FastML Team (2024) https://github.com/fastmachinelearning/hls4ml
Feng B, Zhou Q, Xing J, Yang Q, Chen Y, Deng Z (2024) Attention-empowered transfer learning method for HVAC sensor fault diagnosis in dynamic building environments. Build Environ 250:111148
Finck C, Beagon P, Clauß J et al (2018) Review of applied and tested control possibilities for energy flexibility in buildings—a technical report from IEA EBC Annex 67 energy flexible buildings. Technical Report, International Energy Agency (IEA)
Fu Q, Hu L, Wu H, Hu F, Hu W, Chen J (2018) A SARSA-based adaptive controller for building energy conservation. J Comput Methods Sci Eng 18(2):329–338
Geng Y, Ji W, Xie Y, Lin B, Zhuang W (2022) A sub-sequence clustering method for identifying daily indoor environmental patterns from massive time-series data. Autom Constr 139:104303
Gunay HB, Shi Z (2020) Cluster analysis-based anomaly detection in building automation systems. Energy Build 228:110445
Guo X (2024) Enhancing AI efficiency: the synergy of transformer models and FPGA technology. In: Proceedings of the artificial intelligence and communication technologies (ICAICT 2024), p 49
Han M, May R, Zhang X (2021) Reinforcement learning methodologies for controlling occupant comfort in buildings. In: Data-driven analytics for sustainable buildings and cities: from theory to application. Springer, Singapore, pp 179–205
Heidari A, Peyvastehgar Y, Amanzadegan M (2024) A systematic review of the BIM in construction: from smart building management to interoperability of BIM and AI. Archit Sci Rev 67(3):237–254
Himeur Y, Elnour M, Fadli F, Meskin N, Petri I, Rezgui Y, Bensaali F, Amira A (2023) AI-BIG data analytics for building automation and management systems: a survey, actual challenges and future perspectives. Artif Intell Rev 56(6):4929–5021
Hu W, Wang X, Tan K, Cai Y (2023) Digital twin-enhanced predictive maintenance for indoor climate: a parallel LSTM-autoencoder failure prediction approach. Energy Build 301:113738
Jang J, Han J, Leigh S-B (2022) Prediction of heating energy consumption with operation pattern variables for non-residential buildings using LSTM networks. Energy Build 255:111647
Javed A, Larijani H, Ahmadinia A, Emmanuel R, Mannion M, Gibson D (2016) Design and implementation of a cloud enabled random neural network-based decentralized smart controller with intelligent sensor nodes for hvac. IEEE Internet Things J 4(2):393–403
Jendoubi I, Bouffard F (2023) Multi-agent hierarchical reinforcement learning for energy management. Appl Energy 332:120500
Jeon B-K, Kim E-J (2021) LSTM-based model predictive control for optimal temperature set-point planning. Sustainability 13(2):894
Ji Y, Wang J, Xu J, Fang X, Zhang H (2019) Real-time energy management of a microgrid using deep reinforcement learning. Energies 12(12):2291
Karaiskos P, Munian Y, Martinez-Molina A, Alamaniotis M (2024) Indoor air quality prediction modeling for a naturally ventilated fitness building using RNN-LSTM artificial neural networks. Smart Sustain Built Environ. https://doi.org/10.1108/sasbe-10-2023-0308
Karijadi I, Chou S-Y (2022) A hybrid RF-LSTM based on CEEMDAN for improving the accuracy of building energy consumption prediction. Energy Build 259:111908
Khalil M, Esseghir M, Merghem-Boulahia L (2021) Federated learning for energy-efficient thermal comfort control service in smart buildings. In 2021 IEEE global communications conference (GLOBECOM)

Khan IU, Javaid N, Taylor CJ, Gamage KA, Ma X (2020) Big data analytics based short term load forecasting model for residential buildings in smart grids. In: IEEE INFOCOM 2020-IEEE conference on computer communications workshops (INFOCOM WKSHPS). IEEE, pp 544–549
Kim T-Y, Cho S-B (2019) Predicting residential energy consumption using CNN-LSTM neural networks. Energy 182:72–81
Lee S, Lee J, Ahn Y (2020) LDA-based model for assessing the defect liability system in residential buildings' maintenance phase. J Perform Constr Fac 34(2):04020007
Li Q, Wen Z, Wu Z, Hu S, Wang N, Li Y, Liu X, He B (2021) A survey on federated learning systems: Vision, hype and reality for data privacy and protection. IEEE Trans Knowl Data Eng 35(4):3347–3366
Liu X, Gou Z (2024) Occupant-centric HVAC and window control: a reinforcement learning model for enhancing indoor thermal comfort and energy efficiency. Build Environ 250:111197
Liu G, Yang J, Hao Y, Zhang Y (2018) Big data-informed energy efficiency assessment of china industry sectors based on k-means clustering. J Clean Prod 183:304–314
Liu Y, Kang Y, Zou T, Pu Y, He Y, Ye X, Zhang Y-Q, Yang Q (2024) Vertical federated learning: concepts, advances, and challenges. IEEE Trans Knowl Data Eng. https://doi.org/10.1109/TKDE.2024.3352628
Luo X, Oyedele LO (2021) Forecasting building energy consumption: adaptive long-short term memory neural networks driven by genetic algorithm. Adv Eng Inform 50:101357
L'heureux A, Grolinger K, Elyamany HF, Capretz MA, Challenges and approaches (2017) Machine learning with big data. IEEE Access 5:7776–7797
Masdoua Y, Boukhnifer M, Adjallah KH (2023) Fault tolerant control of HVAC system based on reinforcement learning approach. In: 2023 9th International conference on control, decision and information technologies (CoDIT)
Matsukawa S, Ninagawa C, Morikawa J, Inaba T, Kondo S (2019) LSTM prediction on sudden occurrence of maintenance operation of air-conditioners in real-time pricing adaptive control. In: Artificial neural networks and machine learning—ICANN 2019: text and time series: 28th international conference on artificial neural networks, Munich, Germany, 17–19 September 2019, proceedings, Part IV. Springer, pp 177–189
Mazhar T, Malik MA, Haq I, Rozeela I, Ullah I, Khan MA, Adhikari D, Ben Othman MT, Hamam H (2022) The role of ML, AI and 5G technology in smart energy and smart building management. Electronics 11(23):3960
Melosik M, Naumowicz M, Kropidłowski M, Marszalek W (2022) Remote prototyping of FPGA-based devices in the IoT concept during the Covid-19 pandemic. Electronics 11(9):1497
Miyasawa A, Fujimoto Y, Hayashi Y (2019) Energy disaggregation based on smart metering data via semi-binary nonnegative matrix factorization. Energy Build 183:547–558
Moulla DK, Attipoe D, Mnkandla E, Abran A (2024) Predictive model of energy consumption using machine learning: a case study of residential buildings in South Africa. Sustainability 16(11):4365. https://doi.org/10.3390/su16114365
Naganathan H, Chong WO, Chen X (2016) Building energy modeling (BEM) using clustering algorithms and semi-supervised machine learning approaches. Autom Constr 72:187–194
Ngadiuba J, Loncar V, Pierini M, Summers S, Di Guglielmo G, Duarte J, Harris P, Rankin D, Jindariani S, Liu M et al (2020) Compressing deep neural networks on FPGAS to binary and ternary precision with hls4ml. Mach Learn Sci Technol 2(1):015001
Ngo N-T, Truong N-S, Truong TTH, Pham A-D, Huynh N-T (2024) Implementing a web-based optimized artificial intelligence system with metaheuristic optimization for improving building energy performance. J Asian Archit Build Eng 23(1):264–281
Nguyen VK, Zhang WE, Mahmood A (2021) Semi-supervised intrusive appliance load monitoring in smart energy monitoring system. ACM Trans Sensor Netw (TOSN) 17(3):1–20
Olanrewaju A, Tan W (2022) An artificial neural network analysis of the satisfaction of hospital building maintenance services. IOP Conf Ser Mater Sci Eng 120:012022
Oliosi E, Calzavara G, Ferrari G (2023) On sensor data clustering for machine status monitoring and its application to predictive maintenance. IEEE Sens J. https://doi.org/10.1109/JSEN.2023.3260314
Parhizkar T, Rafieipour E, Parhizkar A (2021) Evaluation and improvement of energy consumption prediction models using principal component analysis based feature reduction. J Clean Prod 279:123866
Patil SR, Sinha MK, Deshmukh MA, Thenmozhi S, Sujatha A (2024) Predicting and forecasting building energy performance using RSM and ANN. Asian J Civ Eng 25(1):159–165
Pekşen MF, Yurtsever U, Uyaroğlu Y (2024) Enhancing for predictive electrical panel anomaly detection maintenance with machine learning and IoI. Alex Eng J 96(87):112–123
Puiu I-L, Fortis T-F (2024) The efficiency of building maintenance using digital twins: a literature review. Adv Inf Netw Appl. https://doi.org/10.1007/978-3-031-57931-8_20
Qin Y, Ke J, Wang B, Filaretov GF (2022) Energy optimization for regional buildings based on distributed reinforcement learning. Sustain Cities Soc 78:103625

Quang TV, Phuong NL (2024) Using deep learning to optimize HVAC systems in residential buildings. J Green Build 19(1):29–50
Raja R, Saraswathi R (2023) An GMM method in IoT approach to improve energy efficiency in smart building. In: 2023 5th International conference on smart systems and inventive technology (ICSSIT)
Ramírez-Sanz JM, Maestro-Prieto J-A, Arnaiz-González A, Bustillo A (2023) Semi-supervised learning for industrial fault detection and diagnosis: a systemic review. ISA Trans 143:255–270
Ren C, Zhu H-C, Wang J, Feng Z, Chen G, Haghighat F, Cao S-J (2023) Intelligent operation, maintenance, and control system for public building: towards infection risk mitigation and energy efficiency. Sustain Cities Soc 93:104533
Roodkoly SH, Fard ZQ, Tahsildoost M, Zomorodian Z (2024) Development of a simulation-based ANN framework for predicting energy consumption metrics: a case study of an office building. Energy Efficiency 17(5):20. 10.1007/s12053-024-10185-1
Rutishauser G (2024) Agile and efficient inference of quantized neural networks PhD thesis, ETH Zurich
Sater RA, Hamza AB (2021) A federated learning approach to anomaly detection in smart buildings. ACM Trans Internet Things 2(4):1–23
Sen S, Yadeo D, Kumar P, Kumar M(2023) Machine learning and predictive control-based energy management system for smart buildings. In: Artificial intelligence and machine learning in smart city planning. Elsevier, Amsterdam, pp 199–220
Seyedzadeh S, Rahimian FP, Glesk I, Roper M (2018) Machine learning for estimation of building energy consumption and performance: a review. Vis Eng 6:1–20
Seyedzadeh S, Rahimian FP, Rastogi P, Glesk I (2019) Tuning machine learning models for prediction of building energy loads. Sustain Cities Soc 47:101484
Sha X, Ma Z, Sethuvenkatraman S, Li W (2023) A new clustering method with an ensemble of weighted distance metrics to discover daily patterns of indoor air quality. J Build Eng 76:107289
Shen R, Zhong S, Wen X, An Q, Zheng R, Li Y, Zhao J (2022) Multi-agent deep reinforcement learning optimization framework for building energy system with renewable energy. Appl Energy 312:118724
Somu N, Mril GR, Ramamritham K (2021) A deep learning framework for building energy consumption forecast. Renew Sustain Energy Rev 137:110591
Song G, Ai Z, Zhang G, Peng Y, Wang W, Yan Y (2022) Using machine learning algorithms to multidimensional analysis of subjective thermal comfort in a library. Build Environ 212:108790
Tian Z, Lu Z, Lu Y, Zhang Q, Lin X, Niu J (2024) An unsupervised data mining-based framework for evaluation and optimization of operation strategy of HVAC system. Energy 291:130043
Tukymbekov D, Saymbetov A, Nurgaliyev M, Kuttybay N, Dosymbetova G, Svanbayev Y (2021) Intelligent autonomous street lighting system based on weather forecast using LSTM. Energy 231:120902
Wang T, Wang C, Zhou X, Chen H (2019) An overview of FPGA based deep learning accelerators: challenges and opportunities. In: 2019 IEEE 21st International conference on high performance computing and communications; IEEE 17th International conference on smart city; IEEE 5th International conference on data science and systems (HPCC/SmartCity/DSS), August. IEEE, pp 1674–1681
Wang JQ, Du Y, Wang J (2020) LSTM based long-term energy consumption prediction with periodicity. Energy 197:117197
Wang C, Wu Z, Peng W, Liu W, Xiong L, Wu T, Yu L, Zhang H (2022a) Adaptive modeling for non-intrusive load monitoring. Int J Electr Power Energy Syst 140:107981
Wang X, Lian L, Yu SX (2022b) Unsupervised selective labeling for more effective semi-supervised learning. In: European conference on computer vision
Wang R et al (2023) Adaptive horizontal federated learning-based demand response baseline load estimation. IEEE Trans Smart Grid. https://doi.org/10.1109/TSG.2023.3318418
Wei Q, Liao Z, Shi G (2020) Generalized actor-critic learning optimal control in smart home energy management. IEEE Trans Ind Inf 17(10):6614–6623
Wen S, Zhang W, Sun Y, Li Z, Huang B, Bian S, Zhao L, Wang Y (2023) An enhanced principal component analysis method with Savitzky–Golay filter and clustering algorithm for sensor fault detection and diagnosis. Appl Energy 337:120862
Wu M-P, Wu F (2024) Predicting residential electricity consumption using CNN-BILSTM-SA neural networks. IEEE Access 12:71555–71565. https://doi.org/10.1109/ACCESS.2024.3400972
Wu H, Huang A, Sutherland JW (2022) Layer-wise relevance propagation for interpreting LSTM-RNN decisions in predictive maintenance. Int J Adv Manuf Technol 118:963–978
Wu R, Ren Y, Tan M, Nie L (2024a) Fault diagnosis of HVAC system with imbalanced data using multi-scale convolution composite neural network. Build Simul 17(371–386):2024. https://doi.org/10.1007/s12273-023-1086-1
Wu Y, Chen S, Jin Y, Xu H, Zhou X, Wang X, Chong A, Li J, Yan D (2024b) Novel occupancy detection method based on convolutional neural network model using PIR sensor and smart meter data. Adv Eng Inform 62:102589

Xu S, Fu Y, Wang Y, O'Neill Z, Zhu Q (2021) Learning-based framework for sensor fault-tolerant building HVAC control with model-assisted learning. In: Proceedings of the 8th ACM international conference on systems for energy-efficient buildings, cities, and transportation

Yanamala AKY (2024) Emerging challenges in cloud computing security: a comprehensive review. Int J Adv Eng Technol Innov 1(4):448–479

Zhang B, Hu W, Ghias AM, Xu X, Chen Z (2022a) Multi-agent deep reinforcement learning-based coordination control for grid-aware multi-buildings. Appl Energy 328:120215

Zhang Z, Mahmud MP, Kouzani AZ (2022b) Resource-constrained FPGA implementation of YOLOv2. Neural Comput Appl 34(19):16989–17006

Zhang L, Chen Z, Ford V (2024a) Advancing building energy modeling with large language models: exploration and case studies. arXiv preprint. arXiv:2402.09579

Zhang M, Millar MA, Chen S, Ren Y, Yu Z, Yu J (2024b) Enhancing hourly heat demand prediction through artificial neural networks: a national level case study. Energy AI 15:100315

Zhang Y, Li N, Zhao T, Li Z (2024c) An energy-saving design method for residential building group based on convolutional neural network. J Build Eng 82:108291

Zhu H, Yang W, Li S, Pang A (2022) An effective fault detection method for HVAC systems using the LSTM-SVDD algorithm. Buildings 12(2):246