Energy & Buildings 367 (2026) 117783

Contents lists available at ScienceDirect

Energy & Buildings

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

A deep-learning framework for predicting building heat load with hyperparameter optimisation and physics-constrained post-processing

Minghui Ma [a], Paolo Valdiserri [a,*], Vincenzo Ballerini [a], Ruixin Li [b], Eugenia Rossi di Schio [a]

[a] Alma Mater Studiorum – University of Bologna, Department of Industrial Engineering DIN, Viale Risorgimento 2, 40136 Bologna, Italy
[b] Alma Mater Studiorum – University of Bologna, Department of Computer Science and Engineering DISI, Viale Zamboni 33, 40126 Bologna, Italy

ARTICLE INFO

Keywords:
Building heat load prediction
Deep learning
Multi-source heterogeneous features
LSTM / TCN / MLP/ XGBoost/LR
Hyperparameter optimisation

ABSTRACT

Driven by global energy transition and carbon neutrality goals, accurate building heat load prediction is of great significance for intelligent heating control, demand-side energy management, and fault detection. This study focuses on a single-apartment located in Bologna and predicts the heat load for the next hour based on the previous 24 h of data. A dataset of 18 typical operating scenarios was constructed, encompassing variations in outdoor meteorological conditions, occupancy, thermostat settings, and equipment heat gains. Subsequently, the Optuna framework was employed to systematically optimize the hyperparameters of five models—LSTM, TCN, MLP, XGBoost, and LR—and their prediction performance was comparatively analyzed. Results show that that the deep learning models LSTM and TCN perform best, with LSTM slightly outperforming (LSTM: MAE = 0.082 kW, RMSE = 0.118 kW, R^2 = 0.863; TCN: MAE = 0.0864 kW, RMSE = 0.1196 kW, R^2 = 0.8606). Whereas the LR model exhibits the poorest performance, achieving only 0.1202 kW, 0.1504 kW, and 0.7793 on the respective metrics. After applying a post-processing (PC) correction mechanism, the MAE and RMSE of LSTM decreased to 0.058 kW and 0.100 kW, corresponding to reductions of 28.2% and 15.3%, respectively. This framework effectively integrates data-driven methods with physical constraints while maintaining low computational cost, providing a scalable and practical solution for intelligent building energy management and low-carbon operation.

1. Introduction

Driven by the global transition toward a low-carbon energy structure and the pursuit of carbon peaking and neutrality goals, energy management in the building sector is facing both critical challenges and transformative opportunities. According to the International Energy Agency (IEA), global building energy consumption exceeded 125 EJ in 2023, accounting for approximately 28% of total final energy use worldwide [1]. Meanwhile, carbon dioxide emissions from building operations and construction peaked at around 10 Gt CO_2 in 2022, contributing 37% of global emissions [2]. With increasing occupant demand for indoor comfort and growing complexity due to climate change, building energy consumption is expected to continue increasing [3]. In this context, improving building energy efficiency is widely recognised as a fundamental strategy for achieving energy conservation and emission reduction targets.

Building heat load, as a critical component of overall building energy consumption, directly impacts heating system efficiency, operational strategies, and economic performance [4]. In intelligent heating systems, the inherent variability of heat load imposes stringent demands on system responsiveness and control precision [5]. Accurate heat load prediction enables proactive energy management, optimizes resource allocation, enhances system performance, and mitigates energy waste, making it an essential technology in smart heating and building energy management [6].

Building heat load prediction can be broadly classified into three categories: white-box models based on physical mechanisms, black-box models driven by data, and hybrid grey-box models that integrate both [7]. White-box models are grounded in thermodynamic and physical laws, offering strong interpretability [8]. However, they are often complex to develop, requiring extensive data on building structure, material properties, and environmental conditions, as well as expert domain knowledge [9]. This complexity leads to high development and maintenance costs, significant computational overhead, and limited applicability to real-time prediction and dynamic control. Grey-box models attempt to integrate physical knowledge with data-driven

* Corresponding author.
E-mail address: paolo.valdiserri@unibo.it (P. Valdiserri).

https://doi.org/10.1016/j.enbuild.2026.117783
Received 4 October 2025; Received in revised form 22 May 2026; Accepted 5 June 2026
Available online 15 June 2026

approaches, achieving a balance between flexibility and reliability [10,11]. Nevertheless, they still face challenges in terms of model complexity, multi-source data fusion, and generalizability [12]. In contrast, black-box models, primarily based on machine learning and deep learning, have gained increasing attention due to their nonlinear modeling capability, high modeling efficiency, strong generalization capability [13]. As a result, they have become the mainstream approach in building heat load prediction research.

Early studies on building heat load prediction primarily relied on traditional statistical modeling approaches, such as multiple linear regression (MLR), autoregressive models (AutoRegressive with eXogenous inputs (ARX), AutoRegressive Moving Average (ARMA)), and recursive least squares (RLS) [14]. For instance, Yun et al. [15] proposed an ARX model based on time and temperature indices for short-term heat load prediction, effectively captured the dominant influencing factors during different time periods and enabled dynamic analysis of heat load variations. Ciulla et al. [16] developed a simplified MLR-based predictive model that, through parameterised simulation and sensitivity analysis, facilitated fast and reliable estimation of heating and cooling demand under varying climatic conditions. Although these statistical models offer simplicity and computational efficiency suitable for basic prediction tasks, they struggle to accurately capture the inherently nonlinear and highly time-varying characteristics of building heat load [17]. Consequently, their accuracy and robustness remain limited, restricting their applicability in high-precision and adaptive real-world systems. This has driven the shift toward more advanced machine learning and deep learning techniques in recent years.

With the rapid development of artificial intelligence, machine learning methods have been widely applied to building heat load prediction due to their advantages in handling nonlinear relationships and high-dimensional data [18]. Representative approaches include artificial neural networks (ANN), random forests (RF), and more recently, gradient boosting models such as XGBoost [19–21]. Muhammad et al. [22] employed ANN and RF to predict hourly HVAC energy consumption in hotel buildings and systematically compared the performance of traditional tree-based models and neural networks under high temporal resolution tasks. However, their study was limited to these two methods, without exploring potentially better-performing models such as support vector regression (SVR) or XGBoost. And their feature sets were relatively simple, failing to fully consider key factors such as weather conditions, equipment efficiency, and occupant behavior. To further improve model performance, some studies introduced optimization algorithms. Satyaki et al. [23] proposed a kernel SVR (kSVR) framework combined with particle swarm optimization (PSO) for energy consumption prediction at the user side of district heating systems. While this approach improved predictive accuracy to some extent, it relied on historical data and had limited ability to capture external disturbances, such as weather variability.

Under complex operating conditions, such shallow models still struggle to capture temporal dependencies and nonlinear dynamics effectively. In contrast, deep learning methods, with their multi-layer nonlinear representation and sequential modeling capabilities, have demonstrated stronger performance in building energy prediction [24]. Among them, long short-term memory (LSTM) is widely used due to their ability to capture long-term dependencies in time series [25]. Wang et al. [26] conducted a systematic comparison of more than twenty ML and deep learning model combinations, showing that LSTM models guided by physical information achieve superior predictive accuracy and stability. Building on this, researchers have further enhanced performance through model fusion. Karijadi and Chou [27] proposed a CEEMDAN-based RF-LSTM hybrid model, in which energy consumption time series were decomposed into different frequency components and predicted separately using RF and LSTM, with the final results integrated to improve accuracy. However, such approaches are generally complex, computationally costly, and highly dependent on the effectiveness of data decomposition, while still insufficiently accounting for external variables such as weather and occupant behavior. Additionally, Zhao et al. [28] proposed a hybrid model combining convolutional neural networks (CNN) with the t-distribution Satin Bowerbird Optimization (tSBO) algorithm for short-term residential heat load prediction. Although optimization algorithms can improve model performance, they also increase training costs, which may limit scalability in large-scale applications.

Model performance depends not only on algorithm structure but also on the construction of input features [29,30]. Building heat load exhibits strong multi-factor interactions, with its variations affected by outdoor weather, heating system operation, indoor setpoints, occupant behavior, and appliance usage patterns [31]. These factors often interact in complex nonlinear and time-varying ways, making feature engineering a crucial determinant of model performance. To address this issue, Ding et al. [32] investigated the effects of different input feature combinations on load prediction, demonstrating that the selection of input variables is crucial for improving model accuracy. However, existing studies still exhibit certain limitations. First, most research tends to rely on static or easily accessible variables (e.g., outdoor temperature), while paying insufficient attention to highly relevant but hard-to-quantify factors such as the dynamic evolution of indoor environments and occupant behavior patterns. Second, the characterization of temporal features remains relatively superficial, lacking systematic modeling of periodic behavioral patterns. These issues limit the generalization ability and predictive accuracy of models in practical applications.

In summary, current building heat load prediction studies still face several key limitations. Existing studies exhibit limitations in both variable selection and operating condition coverage. On the one hand, they tend to rely on static or easily accessible variables (e.g., outdoor temperature), while paying insufficient attention to critical yet difficult-to-quantify factors, such as the dynamic evolution of indoor environments and occupant behavior patterns. On the other hand, most studies are conducted under single operating conditions or limited scenario datasets, making it difficult to fully capture the temporal variability and coupled effects of factors such as occupancy, indoor setpoints, and appliance heat gains. These limitations constrain the ability to accurately represent dynamic load characteristics and undermine their generalization capability and predictive accuracy. Moreover, pure data-driven models often struggle to predict highly variable operating conditions, such as on/off switching or sudden load changes. Improving predictive accuracy typically requires increasing model complexity or employing more sophisticated training strategies, which in turn raises computational costs and reduces practical deployability. In addition, traditional model tuning heavily relies on manual experience and repeated trials, introducing subjectivity and affecting the fairness of model comparison and reproducibility of experiments.

To address these challenges, this study focuses on a single-apartment building in Bologna and develops a heat load prediction framework that balances predictive accuracy, physical plausibility, and computational cost. Specifically, the main contributions of this work are:

1. A high-quality multi-source dynamic heat load dataset covering 18 typical operating scenarios was constructed, systematically considering occupancy, indoor setpoints, and appliance thermal gains. The dataset not only captures the building thermal dynamics comprehensively but also includes a substantial number of zero-load samples, reflecting realistic equipment on/off cycling and providing a solid foundation for model training.
2. A physical constraint post-processing (PC) mechanism is proposed to correct predictions based on boiler operation states, ensuring that outputs satisfy physical boundaries without additional training cost, thereby enhancing reliability under extreme conditions.
3. An automated hyperparameter optimization framework based on Optuna, using the TPE algorithm, was established to tune multiple models consistently, reducing manual intervention and improving the fairness and reproducibility of model comparisons.

4. Five representative models—linear regression (LR), XGBoost, MLP, LSTM, and temporal convolutional network (TCN)—were systematically compared to evaluate predictive performance under multiple scenarios, providing a reliable basis for model selection and optimization in building heat load prediction.

Nomenclature

Abbreviations

ACF	autocorrelation function
ANN	artificial neural network
ARX	autoRegressive with EXogenous inputs
ASHP	air source heat pump
BPNN	backpropagation neural network
CCF	cross-correlation function
CEEMDAN	complete ensemble empirical mode decomposition with adaptive noise
CI	confidence intervals
COP	coefficient of performance
ELM	extreme learning machine
HVAC	heating, ventilation, and air conditioning
IEA	International Energy Agency
IQR	interquartile range
LightGBM	light gradient boosting machine
LR	linear regression
LSTM	long short-term memory
MAE	mean absolute error
MLP	multilayer perceptron
MLR	multiple linear regression
MSE	mean squared error
OAT	one-at-a-Time
OLS	ordinary least squares
PINNS	physics-informed neural networks
R^2	coefficient of determination
RF	random forest
RL	reinforcement learning
RLS	recursive least squares
RMSE	root mean squared error
SVM	support vector machine
SVR	support vector regression
TCN	tabular temporal convolutiona network
TPE	tree-structured Parzen estimator
tSBO	T-distributed Satin Bowerbird
XGBoost	extreme gradient boosting

2. Methodology

2.1. Data acquisition and preprocessing

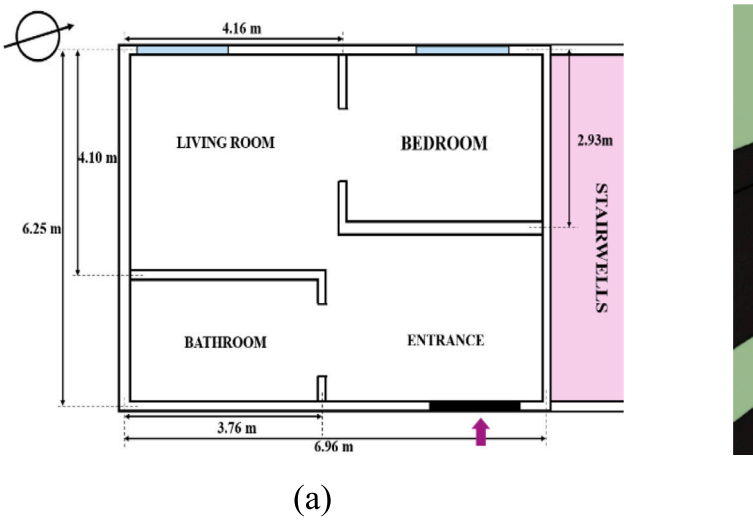
2.1.1. Data acquisition

This study focuses on a residential apartment located on the ground floor of a four-story building in Bologna, Italy, with a total floor area of approximately 40 m^2. The internal layout comprises a bedroom, bathroom, living room, and entrance area, as illustrated in Fig. 1. The building envelope consists of horizontally perforated bricks with integrated insulation, and the apartment is equipped with two external windows. The dimensions and thermophysical properties of the main envelope components are summarized in Table 1.

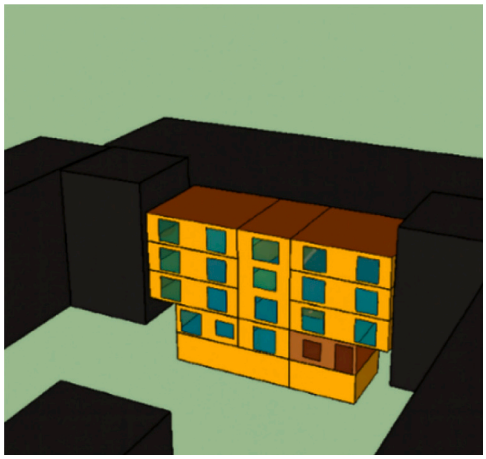
According to local regulations, the official heating season in Bologna extends from October 15 to April 15 of the following year. To generate heat load data for model development, this study extends a previously validated 3D building model and TRNSYS-based dynamic simulation framework [33,34], conducting and performs systematic simulations under multivariate conditions. Unlike the original study, which focused on thermal comfort and CFD analysis, the simulation scheme in this work was redesigned to generate a multi-source dataset for heat load prediction through the construction of multiple operational scenarios. To comprehensively capture the dynamic characteristics of heat load, several key influencing factors were selected as simulation input (see Table 2), including outdoor meteorological conditions, indoor temperature control strategies, and internal heat gains. Hourly meteorological data for Bologna, including outdoor air temperature and relative humidity, were obtained from the Open-Meteo data platform. Three daytime temperature setpoints (06:00–23:00) were defined: 20.5 °C,

Table 1
Transmittances U and apartment envelope components' dimensions.

Component	Thickness (m)	Area (m^2)	U (W/m^2K))
External walls	0.3	–	0.667
Dividing walls	0.1	–	2.047
Inter-floor	0.42	–	0.595
Entrance door	–	2.43	5.54
Window(bedroom)	–	1.89	1.69
Window (living room)	–	5.94	1.69



(a)



(b)

Fig. 1. Layout of the apartment (a) and a 3D view of the building modelled with Google SketchUp, where the apartment analyzed is shown in light brown (b). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Summary of influencing factors.

Variables	Description
Temperature	Open-Meteo
Relative Humidity	Open-Meteo
Indoor temperature set-points	20.5 °C (6:00—23:00) 17.5 °C (23:00–6:00)
	21.5 °C (6:00—23:00) 17.5 °C (23:00–6:00)
	22.5 °C (6:00—23:00) 17.5 °C (23:00–6:00)
Occupant behavior	Single occupancy
	Double occupancy
Electrical appliances	According to IEA Task 44 for dynamic analyses [35]

21.5 °C, and 22.5 °C and a unified nighttime setpoint of 17.5 °C (23:00–06:00). Internal gains were modelled according to IEA Task 44 [35], considering two typical occupancy scenarios (single and double occupancy) to reflect variations in heat due to different numbers of residents. In addition, a ± 20% perturbation was introduced to the baseline heat gains from standard electrical appliances to evaluate the impact of appliance load fluctuations on heat load. The simulation period covers the entire heating season from 15 October 2022 to 15 April 2023, with a temporal resolution of 1 h. A full-factorial design was adopted to generate different operational scenarios, resulting in a total of 18 representative simulation scenarios (i.e., 2 occupancy conditions × 3 temperature setpoints × 3 appliance heat-gain levels). Since heat load is typically low at the beginning and end of the heating season due to relatively higher outdoor temperatures, this period provides limited representative heat load data. To ensure the validity and representativeness of the analysis, the simulation dataset was filtered to retain periods with stable heat load, covering the early, middle, and late stages of the heating season. Ultimately, simulation results from 30 October 2022 to 25 March 2023 were selected for subsequent analysis. Fig. 2(a) presents the variation in the daily average heat load of the building during this period, while Fig. 2(b) shows the hourly heating demand for a typical week from 15 January to 21 January 2023. As illustrated in the figures, the building heat load exhibits a clear diurnal cycle, primarily influenced by variations in outdoor temperature as well as the daily activity patterns of occupants.

2.1.2. Data preprocessing

To ensure data quality and model stability, a systematic preprocessing workflow was implemented, mainly included data distribution analysis, outlier detection and filtering, and feature normalization. First, the overall distribution characteristics of the data were analyzed, with the distribution of heat loads shown in Fig. 3(a). Outlier detection was then performed using the interquartile range (IQR) method combined with boxplots Fig. 3(b). To handle the detected outliers, further analysis was conducted in conjunction with meteorological variable variations and the temporal continuity of the time series to distinguish their causes. The isolated spikes or drops that were not accompanied by significant extreme weather conditions were classified as simulation artifacts caused. In these cases, the original values were corrected via interpolation using adjacent time-step data to preserve the continuity of the time series. Conversely, outliers aligned with extreme outdoor temperatures or other notable meteorological changes were considered to reflect genuine extreme operating conditions and were therefore retained. After this screening process, 63504 valid samples were obtained. Among them, 11630 samples correspond to zero heating load, accounting for 18.31% of the dataset. The dataset was divided into three subsets: a training set, a validation set, and a test set. Specifically, 70% of the data (from October 30, 2022 to February 15, 2023) were used for model training, 15% (from February 16, 2023 to March 5, 2023) for validation, and the remaining 15% (from March 6, 2023 to March 25, 2023) were reserved as an independent test set to evaluate the final predictive performance of the model. Additionally, because input variables differ significantly in physical dimension and value range, Max–Min normalization was applied to linearly scale all features to the [0,1] range. Crucially, to avoid data leakage, the maximum and minimum values used for normalization were derived exclusively from the training set and then applied to the test set. This procedure ensures a rigorous and unbiased evaluation of model performance on previously unseen data.

2.2. Model development

2.2.1. Input variable selection and feature engineering

Based on the high-quality simulation data, a set of multi-dimensional input features was constructed to characterize the dynamic variations of building heat load. The features include meteorological conditions, internal heat gains, system control states, historical load, and temporal characteristics. Together, these features capture the primary physical and operational factors affecting building heat load. Beyond instantaneous variables, building heat load exhibits significant temporal dependence, due to the thermal inertia of the building envelope and the periodic operation of the heating system. To characterize these dynamic features, historical lag features are introduced. First, the autocorrelation function (ACF) of the target variable P_{th} was analyzed (Fig. 4(a)). The results indicate a pronounced correlation peak around 24 h, demonstrating the strong daily periodicity of building heat load. This observation is consistent with the hourly heat load patterns shown in Fig. 2 (b). Moreover, the ACF remains relatively high at short lags, indicating that current load is strongly influenced by the preceding time steps. Based on this analysis, short-term thermal inertia is captured using $P_{th,lag1}$ and $P_{th,lag2}$, while daily periodic operation is represented by $P_{th,lag24}$ and $P_{th,lag48}$. Furthermore, the cross-correlation function (CCF) between building heat load and outdoor temperature was analyzed (Fig. 4(b)). The results show a significant negative correlation, with the peaks

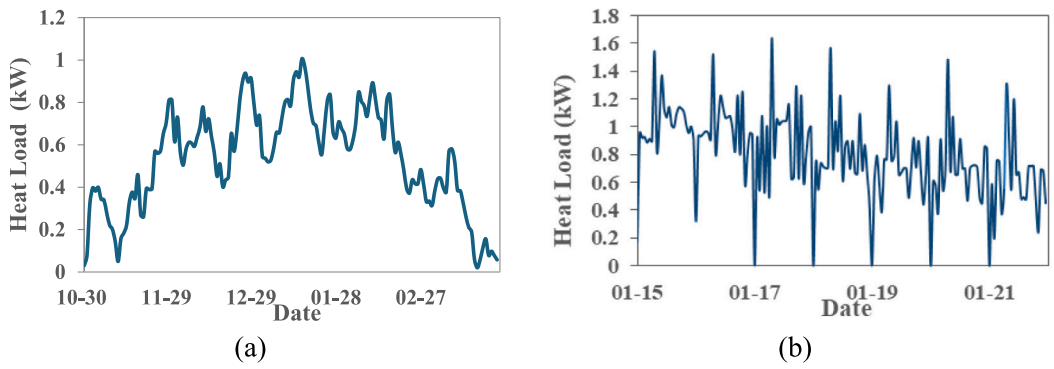


Fig. 2. Distribution of building heat load during the heating season: (a) Variation of daily average heat load (Oct 30, 2022–Mar 25, 2023); (b) Typical weekly hourly heat load profiles (Jan 15–Jan 21, 2023).

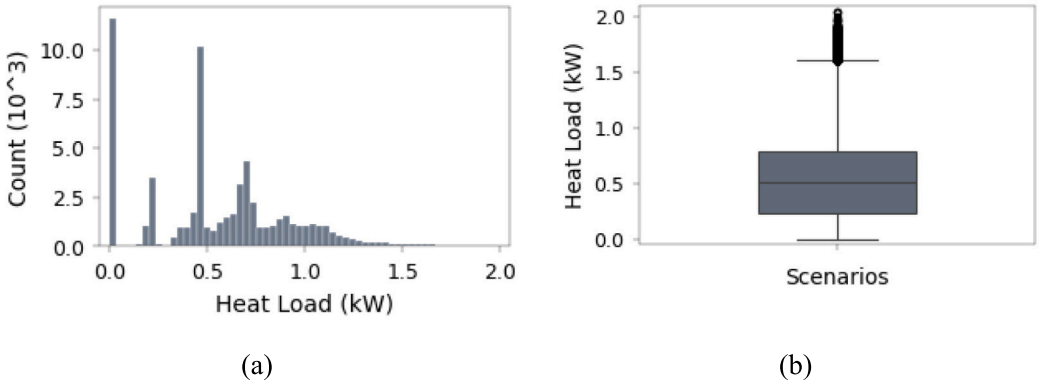


Fig. 3. Statistical distribution and outlier analysis of heat load data. (a) Distribution histogram of heat load counts. (b) Boxplot for outlier detection.

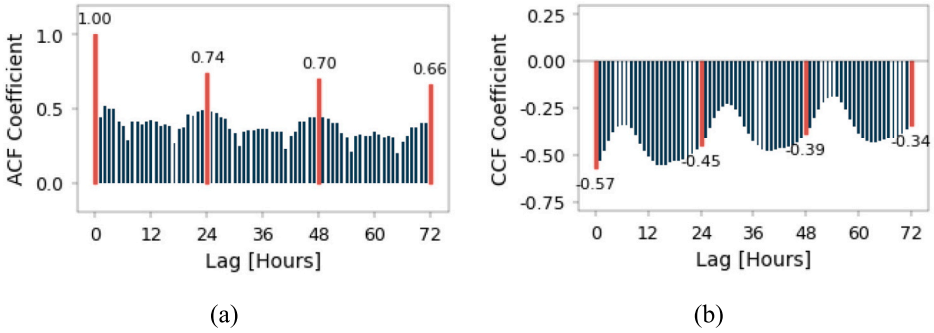


Fig. 4. ACF and CCF analysis of building heat load: (a) Autocorrelation of heat load; (b) Cross-correlation between outdoor temperature and heat load.

recurring every 24 h, reflecting the influence of diurnal cycles. At lag0, the CCF reaches 0.57 (strongest correlation), and the absolute value gradually decreases with increasing lag, indicating the influence of outdoor temperature on heating load weakens over time. Based on this analysis, short-term lag features for outdoor temperature $T_{out,lag1}$ and $T_{out,lag2}$, as well as the daily periodic feature $T_{out,lag24}$ were introduced to capture the delayed response of heat load to weather variations. Initial exploration also considered including lagged outdoor relative humidity features, but feature importance analysis indicated negligible contribution to prediction performance. To maintain model simplicity and emphasize the primary physical drivers, relative humidity lag features were excluded.

In addition, building heat load exhibits a pronounced daily cycle. To capture this periodicity, cyclical encoding using sine and cosine functions was applied to hourly information, generating the features h_{cos} and h_{sin}. This approach avoids discontinuous jumps in time representation (e.g., between 23:00 and 00:00), enabling the model to more accurately learn the periodic patterns of heat load. To further capture seasonal variations within the heating season, the temporal information of the heating season was also cyclically encoded. Specifically, the starting date (30 October 2022) was used as a reference, and the number of days from this date was calculated for each time step to generate the seasonal time variable. Considering the heating season lasts 154 days, sine and cosine transformations were applied to construct feature s_{cos} and s_{sin} Table 3 presents a summary of the final input features used in this study.

Table 3
Input features.

Category	Feature	Description
Outdoor environment	T_{out} (°C)	Outdoor air temperature
	$T_{out,lag1}$, $T_{out,lag2}$, …$T_{out,lag24}$ (°C)	Outdoor temperature 1, 2, 24 h before
	H_{out} (%)	Outdoor relative humidity
Thermal comfort	$T_{setpoint}$ (°C)	Indoor temperature setpoint
Internal gains	Equipment (W/m^2)	Equipment power density
	Occupancy (Person)	Number of occupants
System operation	Boiler on (−)	Boiler operating state (0/1)
	Day night (−)	Day-night operation indicator
Historical load	$P_{th,lag1}$, $P_{th,lag2}$, $P_{th,lag24}$, $P_{th,lag48}$ (kW)	Heat load 1, 2, 24, 48 h before
Time features	h_{sin}, h_{cos}	Hourly cyclic encoding
	s_{sin}, s_{cos}	Seasonal cyclic encoding

2.2.2. Algorithm selection

This study aims to develop and systematically compare multiple indoor models for heat load prediction to identify the most effective approach. To achieve this objective, five representative algorithms are selected based on prior research and their characteristics, covering classical regression, ensemble learning, and deep learning methods, thereby enabling a comprehensive evaluation of their applicability and predictive performance in heat load prediction.

LR, as classical statistical method, is employed to establish a linear mapping between input features (e.g., outdoor temperature and humidity) and indoor heat load using the ordinary least squares (OLS). Due to its simplicity and strong interpretability, LR is commonly adopted as a baseline model to benchmark the performance of more advanced methods.

XGBoost is an ensemble learning method based on gradient boosting decision trees. It iteratively constructs weak learners and optimizes the loss function along the negative gradient direction, while incorporating regularization to control model complexity and mitigate overfitting. In indoor heat load prediction, XGBoost can effectively handle high-dimensional input features (e.g., meteorological variables and operational patterns) and capture complex nonlinear relationships, resulting in strong predictive accuracy and generalization performance. Consequently, it has become one of the most widely used ensemble learning

methods in this field.

MLP is a typical feedforward neural network consisting of an input layer, multiple hidden layers, and an output layer. By applying nonlinear activation functions, MLP performs hierarchical feature transformations, enabling it to model complex nonlinear relationships between inputs and heat load. It is considered a fundamental deep learning model for regression tasks.

LSTM is an improved variant of recurrent neural networks (RNN) that introduces memory cells along with input, forget, and output gates to selectively retain and update information, thereby alleviating the vanishing gradient problem. In indoor heat load prediction, LSTM is capable of capturing both short-term fluctuations and long-term dependencies in time series data, making it one of the most used deep learning models for heat load prediction.

TCN is a convolution-based sequence modeling approach that employs causal and dilated convolutions to process time series data in parallel. By enlarging the receptive field while maintaining computational efficiency, TCN can effectively extract multi-scale temporal features and capture both short-term variations and long-term dependencies. In heat load prediction tasks, TCN typically demonstrates stable performance, showing strong application potential.

In summary, this study systematically compares the above five models under a unified framework to evaluate their performance in indoor heat load prediction, providing a basis for selecting high-accuracy prediction methods.

2.3. Hyperparameter optimization and training configuration based on Optuna

To mitigate subjective bias in manual hyperparameter tuning and ensure a fair comparison of model performance, this study develops an automated hyperparameter optimization framework based on Optuna. Compared to traditional grid or random search, Optuna employs the TPE Bayesian optimization algorithm, which guides the search process by leveraging the prior distribution of historical samples. By prioritizing high-potential regions within complex high-dimensional parameter spaces, this strategy enhances computational efficiency and reduces redundant trials, making it particularly well-suited for heat load prediction tasks characterized by long training cycles and expansive hyperparameter spaces. All models were trained on an NVIDIA L4 GPU. Based on ACF analysis, the historical observation window was uniformly set to 24-time steps. Sequential models such as LSTM and TCN directly receive input tensors of dimension 24 times features. In contrast, non-sequential models utilize a flattening technique to map time-step features into a unified vector space, ensuring that all models maintain equivalent input information entropy. For the training configuration, deep learning architectures were trained utilizing the Adam optimizer with a dropout rate of 0.2 and a batch size of 32. An early stopping mechanism is implemented to suppress overfitting, whereby training is terminated if the validation loss fails to improve for 10 consecutive epochs (with a maximum limit of 100 epochs). Each model underwent 50 optimization trials. The structured hyperparameter search space is detailed in Table 4, ensuring that all algorithms are evaluated under their respective optimal configurations to guarantee experimental reproducibility and scientific rigor.

2.4. Evaluation indices

In this study, three evaluation metrics are employed to comprehensively assess the performance of the prediction models from multiple perspectives, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2). MAE and RMSE measure the magnitude of prediction errors, with lower values indicating higher predictive accuracy. The R^2 reflects the model's goodness of fit, with values ranging from 0 to 1; a value closer to 1 signifies a better fit between the predicted and actual values.

Table 4
Hyperparameter search space for each prediction model.

Algorithm	Hyperparameter	Search space
LSTM	Layers	{1,2}
	Hidden units	[16, 64], step = 16
	Learning rate	[1e-4, 1e-1] (log scale)
TCN	Filters	[16, 64], step = 16
	Kernel size	{3, 5}
	Blocks	[2,4]
	Dilations	2^n, n ∈ [0, blocks-1]
	Learning rate	[1e-4, 1e-2] (log scale)
MLP	Hidden units	[16, 64], step = 16
	Layers	[2,4]
	Activation	{relu, tanh}
	Learning rate	[1e-4, 1e-1] (log scale)
XGBoost	n_estimators	[100, 600], step = 50
	max_depth	[3,5]
	learning_rate	[0.005, 0.1]
	reg_alpha	[1e-3, 10] (log scale)
	reg_lambda	[1e-3, 10] (log scale)
	gamma	[0.1, 5.0] (log scale)
	subsample	[0.5, 0.9]
	colsample_bytree	[0.5, 0.9]
	min_child_weight	[1,15]

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

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

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{3}$$

wherey_i represents the actual values, $\hat{y}_i$represents the predicted values, and n is the number of observations.

3. Results and discussion

3.1. Model comparison

Table 5 summarizes the optimal hyperparameters for each model obtained through the Optuna-based automated optimization framework. The LSTM model, the optimal architecture is a single-layer network with 64 hidden units and a learning rate of 0.001. The TCN

Table 5
Optimal hyperparameter configurations obtained via Optuna.

Algorithm	Hyperparameter	Value
LSTM	Layers	1
	Hidden units	64
	Learning rate	0.001
TCN	Filters	32
	Kernel size	5
	Blocks	4
	Dilations	{1, 2, 4, 8}
	Learning rate	0.00028
MLP	Hidden units	64
	Layers	4
	Activation	tanh
	Learning rate	0.00052
XGBoost	n_estimators	350
	max_depth	5
	learning_rate	0.069
	reg_alpha	0.094
	reg_lambda	0.13
	gamma	0.13
	subsample	0.89
	colsample_bytree	0.66
	min_child_weight	14

model is configured with 32 convolutional filters, a kernel size of 5, and 4 residual blocks with dilation factors of {1, 2, 4, 8}, using a learning rate of 0.00028. The MLP model adopts a 4 layer architecture, with each layer containing 64 hidden units, employing the tanh activation function and a learning rate is 0.00052. For the XGBoost model, the optimal parameters are 350 base learners, a maximum tree depth of 5, and a learning rate of 0.069. The regularization parameters are $\alpha = 0.094$ and $\lambda = 0.13$, $\gamma = 0.13$, a subsample ratio of 0.89, a column sampling ratio of 0.66, and a minimum child weight is 14.

Table 6 systematically summarizes the quantitative evaluation results of the five heat load prediction models, including MAE, RMSE, R^2, 95% CI, and training time. Fig. 5 provides a visual comparison of model performance, illustrating the differences in MAE, RMSE, and R^2 across the models. The results indicate that the deep learning models—LSTM and TCN—achieve the highest predictive. Specifically, the LSTM model yields MAE, RMSE, and R^2 values of 0.082 kW, 0.118 kW, and 0.863, respectively, while the TCN model achieves 0.0864 kW, 0.1196 kW, and 0.8606. The performance is highly comparable, suggesting that both models are capable of effectively capturing the temporal dependencies in heat load data. From a physical perspective, the dynamic behavior of building heat load is governed by the thermal inertia of the building envelope, complex heat transfer processes, and the lag effect between external disturbances and indoor thermal response. These factors lead to pronounced temporal correlations and delayed characteristics in the heat load series. The thermal response characteristics of a building can be represented by the thermal time constant, $\tau = C / H$, where C denotes the effective thermal capacity of the building and H represents the total heat loss coefficient. Based on the estimated parameters of the Bologna apartment, the theoretical thermal time constant was calculated to be approximately 42.19 h. The complete calculation procedure is provided in Appendix A. The results indicate that the building possesses strong thermal inertia and significant thermal lag effects. The LSTM model, with its gated memory mechanism, capture long-term dependencies and retain historical thermal state information. In contrast, the TCN model employs dilated causal convolutions to expand the receptive field and capture multi-scale temporal features and lag effects. Therefore, the superior predictive performance of both LSTM and TCN fundamentally arises from their ability to effectively learn and represent the intrinsic thermodynamic evolution and temporal characteristics of heat load systems. MLP model ranks third in predictive performance, with MAE, RMSE, and R^2 values of 0.0910 kW, 0.1273 kW, and 0.842, respectively, indicating relatively robust overall performance. This can be primarily attributed to the short-term forecasting task considered in this study, where the heat load of the next hour is predicted based on the previous 24 h of data, thereby reducing the complexity of temporal dependencies. In contrast, XGBoost and LR exhibit weaker predictive performance. LR performs the worst (MAE = 0.1202 kW, RSME = 0.1504 kW, $R^2 = 0.7793$), highlighting the limitations of traditional linear models in capturing the complex nonlinear characteristics of heat load data. Furthermore, the CI of LSTM ([0.0806, 0.0842] kW,) and TCN ([0.0847, 0.0882] kW) are notably narrower, indicating that both models demonstrate greater stability and reliability in handling load fluctuations. Computational costs of the models differ significantly. LSTM and TCN are the longest (463.22 s and 469.08 s), MLP is intermediate (158.41 s), and XGBoost and LR are the shortest (47.68 s and 5.09 s), all well below one hour. LSTM and TCN exhibit higher computational costs due to the need to process sequential data through extensive recurrent or convolutional operations, whereas MLP, despite its multi-layered architecture, entails lower computational complexity; in contrast, XGBoost and LR rely on tree-based structures or linear computations, resulting in greater computational efficiency. In summary, for heat load prediction scenarios that demand high prediction accuracy and stability, LSTM and TCN are the preferred choices. Conversely, in applications with limited computational resources or strict requirements on training efficiency, MLP serves as a viable alternative.

Fig. 6 illustrates the predicted heat load curves of different models over the test period (March 6–25, 2023). To more clearly characterize model performance along the temporal dimension, the test period is divided into three sub-phases: (a) March 6–12, (b) March 13–19, and (c) March 20–25.

As shown in Fig. 6, the predicted curves of LSTM (purple) and TCN (orange) closely follow the true values (black solid line), indicating strong fitting capability. Notably, during periods of pronounced nonlinear fluctuations, both models effectively capture the rising and falling edges of load changes. This observation is consistent with the quantitative results presented in Table 6, confirming their capability in modeling heat load time series with long-term dependencies. To evaluate model robustness under critical conditions, both peak-tracking capability and low-load stability are analyzed. Regarding peak tracking, instantaneous load peaks occurred around 11:00 on March 8 and March 9. LSTM and TCN respond rapidly to these changes, accurately capturing increases in heat load. In contrast, during the low-load period after March 14, the actual heat load frequently approached zero. During this phase, the predictions of the MLP (green) and XGBoost (red) showed noticeable oscillations and high uncertainty, while LSTM and TCN produced smoother predictions that closely followed the baseline. LR (blue) exhibited significant prediction failures in subplots (b) and (c). Specifically, around 12:00 on March 17 and March 24, LR predictions deviated substantially from the actual values, even producing physically implausible negative values (below −0.1 kW). This indicates the limitations of linear models in capturing complex, non-stationary temporal relationships. In terms of smoothness and consistency with the actual data distribution, LSTM and TCN show smaller fluctuations and stronger signal reconstruction capability. This aligns with the narrower 95% CI reported in Table 6, indicating higher robustness in handling external disturbances and random fluctuations. In contrast, XGBoost and LR are more prone to deviate from the true trajectory over long time-series predictions, suggesting that traditional regression models are more susceptible to cumulative errors in prediction, leading to reduced generalization performance. Overall, the fitting results in Fig. 6 LSTM and TCN outperform other models in heat load prediction, providing stable and reliable prediction under complex dynamic conditions.

To evaluate the predictive performance of each model from a statistical distribution perspective, Fig. 7 presents scatter plots comparing predicted and actual values for the five algorithms on the test set. The clustering of points around the reference line Y = X provides an intuitive measure of model fitting capability. The scatter points of LSTM and TCN are the most compact, distributed almost symmetrically around the diagonal. This aligns with the quantitative results in Table 6, indicating that deep learning models achieve the highest accuracy in heat load prediction. For MLP, the scatter shows moderate deviation, expanding

Table 6
Model performance evaluation results.

Algorithms	MAE (kW)	RMSE (kW)	R^2	95%CI (MAE (kW))	95%CI (RMSE(kW))	Training time(S)
LSTM	0.082	0.118	0.863	[0.0806, 0.0842]	[0.1160, 0.1207]	463.22
TCN	0.0864	0.1196	0.8606	[0.0847, 0.0882]	[0.1173, 0.1218]	469.08
MLP	0.0910	0.1273	0.842	[0.0892, 0.0928]	[0.1250, 0.1295]	158.41
XGBoost	0.1026	0.1328	0.828	[0.1010, 0.1043]	[0.1308, 0.1347]	47.68
LR	0.1202	0.1504	0.7793	[0.1183, 0.1220]	[0.1481, 0.1528]	5.09

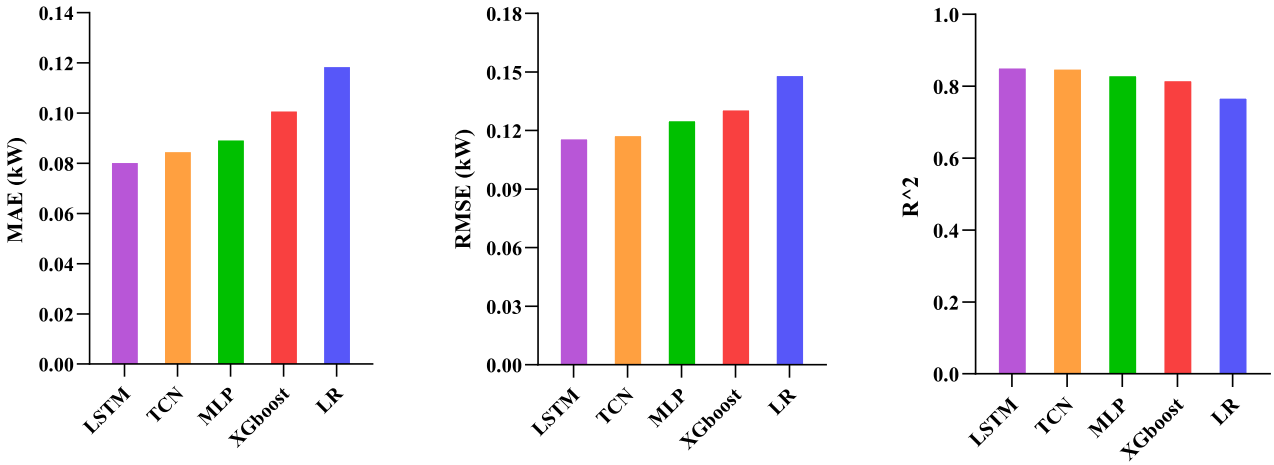


Fig. 5. Quantitative comparison performance of the prediction models using MAE, RMSE, and R^2.

outward in the medium-load range (0.6–1.2 kW), reflecting sensitivity to data fluctuations and random oscillations. XGBoost exhibits dense downward deviations near the low-load region (around 0 kW), suggesting limited robustness of tree-based models to extreme values and noise. LR shows the most dispersed scatter, forming a fan-like pattern that deviates furthest from the reference line, highlighting its inability to capture complex nonlinear relationships and the lowest prediction stability.

In the analysis of critical operating conditions, vertical deviations in the zero-load region provide an intuitive measure. LR exhibits the largest deviation, with predicted values ranging from −0.8 to 0.5 kW and producing numerous negative predictions. This "vertical band" pattern indicates that LR fails to accurately identify equipment shutdown states. XGBoost shows moderate scatter expansion near zero. Although its nonlinear modeling capability exceeds that of LR, the lack of temporal memory limits its ability to recognize shutdown sequences, resulting in higher baseline noise and elevated MAE and RMSE. In contrast, deep learning models demonstrate higher prediction stability under shutdowns. LSTM model effectively retains memory of the shutdown state through its gating and forget mechanisms, reducing scatter fluctuations. TCN model leverages dilated convolutions to capture long-range temporal dependencies, improving the accuracy of operating state transitions. Overall, sequence-based deep learning methods more accurately reflect the physical state of equipment in the zero-power region, significantly outperforming non-temporal models. Considering both scatter distributions and critical operating condition analysis, LSTM and TCN achieve the best overall performance in terms of predictive accuracy and shutdown state recognition, followed by MLP. XGBoost and LR exhibit notable deviations under low-load and shutdown conditions, highlighting the limitations of traditional non-sequential methods in complex heat load prediction.

3.2. *Robustness analysis under noisy conditions*

To comprehensively evaluate the robustness and noise resilience of the models under complex conditions, two types of representative stochastic noise—Gaussian noise and Laplacian noise—were introduced into the original dataset to simulate different forms of uncertainty. Gaussian noise, following a normal distribution, is commonly used to characterize pervasive random errors and thermal noise in real systems. It simulates uniform disturbances caused by sensor measurement limitations and signal transmission inaccuracies, enabling the assessment of model stability under typical random perturbations. In contrast, Laplacian noise exhibits pronounced heavy-tailed characteristics and can emulate abrupt or impulsive disturbances, such as extreme weather events or anomalous data injections. Introducing Laplacian noise thus provides a means to further examine robustness and fault-tolerance in the presence of outliers.

In this study, noise intensities ranging from 0 to 0.2 were introduced to systematically evaluate model robustness under varying perturbation levels. Specifically, noise levels of 0.05, 0.1, and 0.2 correspond to low, medium, and high noise, simulating minor measurement errors, transmission disturbances, and extreme environmental perturbations, respectively. Tables 7 and 8 summarize the predictive performance of the models under Gaussian and Laplacian noise. Fig. 8 provides a visual comparison, illustrate the impact of noise type and magnitude on model performance. The left column (a, c, e) represents Gaussian noise scenarios, whereas the right column (b, d, f) corresponds to Laplacian noise. Overall, as the noise level increases, the performance of all models deteriorates, as reflected by increasing MAE and RMSE and decreasing R^2, indicating the negative impact of noise on predictive accuracy. Among all models, LSTM and TCN demonstrate the strongest noise resilience across all noise levels. Under Gaussian noise, the MAE of LSTM increases marginally from 0.0900 to 0.0951 kW, while its R^2 decreases slightly from 0.8537 to 0.8346. TCN exhibits a highly similar trend and, even achieves slightly better R^2 values under certain high-noise conditions. A similar pattern o is observed under Laplacian noise, further confirming the superior noise resilience of these two models. The superior robustness of LSTM and TCN can be attributed to their structural design. LSTM leverages input, forget, and output gates along with memory cells to selectively integrate temporal information and accumulate stable patterns, effectively suppressing the propagation of high-frequency noise. Similarly, TCN employs causal convolutions with a well-defined temporal receptive field, allowing the model to smooth input perturbations. These architectural features allow both models to maintain stable temporal representations and achieve predictive robustness under noisy conditions. MLP exhibits a degradation trend similar to that of LSTM and TCN across all noise levels, indicating a comparable response to noise, although its overall prediction errors remain higher. Compared with neural network models, XGBoost demonstrates higher sensitivity to noise. When only 5% Gaussian noise is introduced, its R^2 drops sharply from 0.8280 to 0.7971, a decrease of approximately 3.7%, which is substantially larger than the corresponding change for LSTM (0.07%). As noise intensity increases to 20%, its MAE and RMSE rise to 0.1137 and 0.1502 kW, respectively. Under Laplacian noise, XGBoost exhibits a similarly increasing trend in errors as noise intensity rises, further indicating that the model is less robust under noisy conditions. From an algorithmic perspective,XGBoost is an additive model that iteratively fits residuals using gradient boosting. During training, the model improves its fitting accuracy by sequentially learning the residuals from the previous iteration. However, when input data contains random noise, the proportion of noise in the residuals gradually increases, increasing

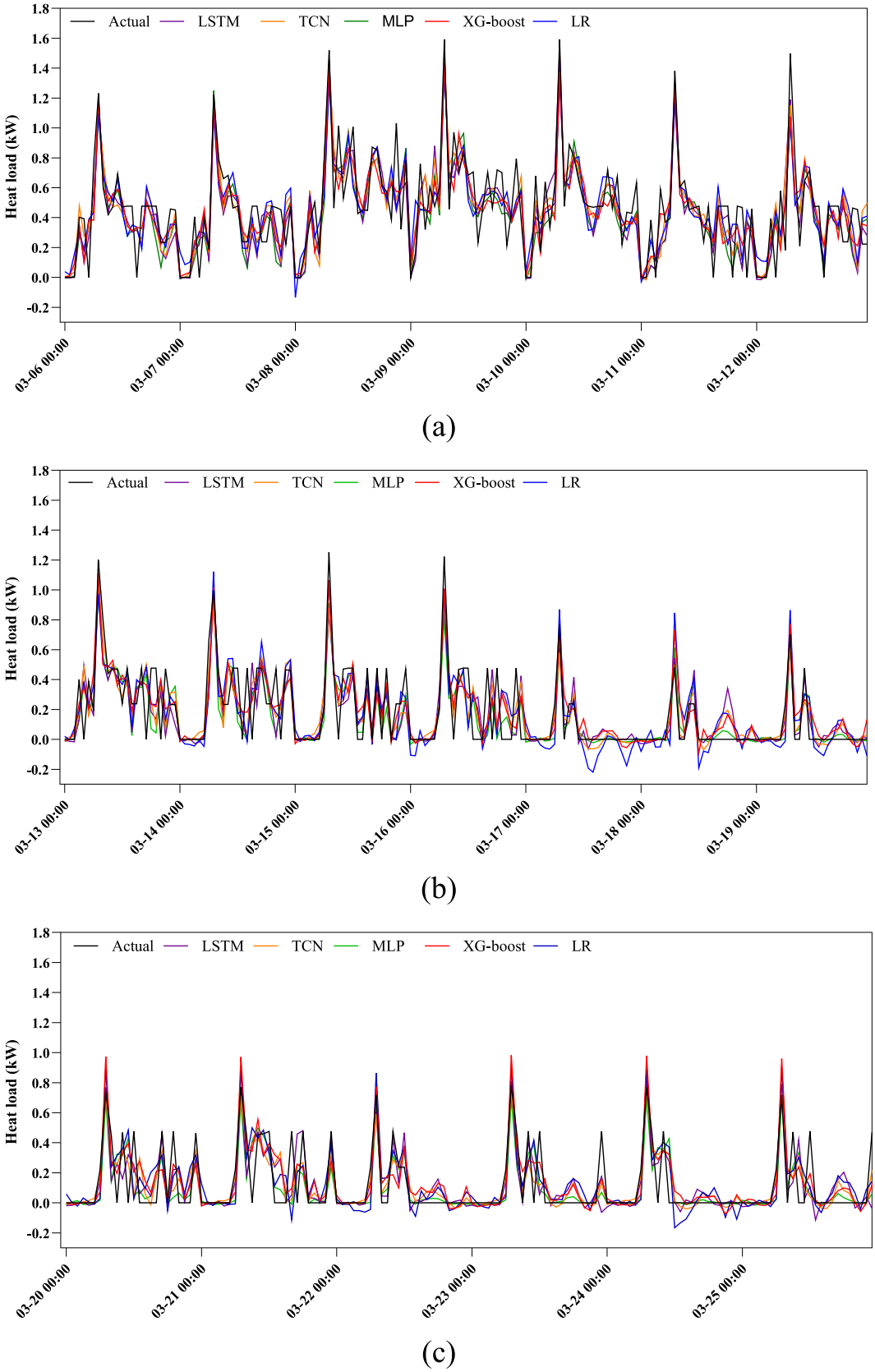


Fig. 6. Heat load prediction curves of different models over the test period (March 6–25, 2023), divided into three sub-phases: (a) March 6–12; (b) March 13–19; (c) March 20–25.

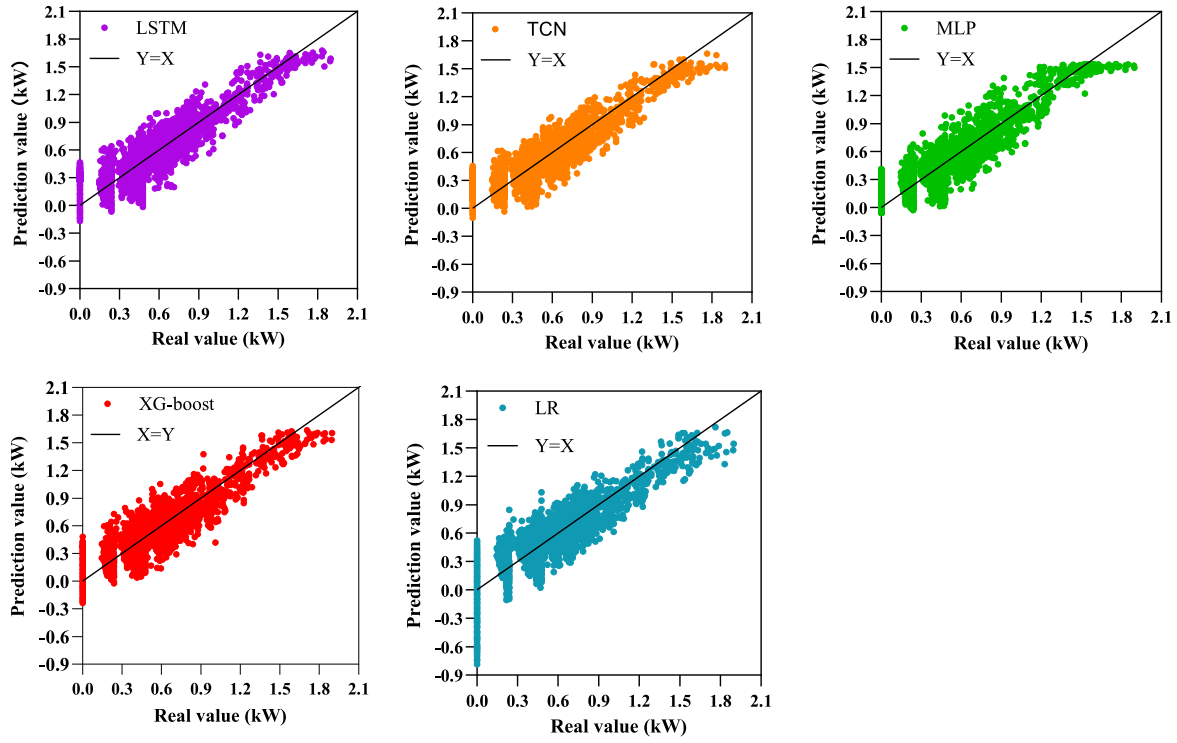


Fig. 7. *Predicted* vs. *actual values* over the test period (March 6–25, 2023). The Y = X black solid line indicates a perfect match between predicted and measured values.

Table 7
Comparison of model predictive performance under Gaussian noise.

Model	Metric	0%(Baseline)	5% Noise	10% Noise	20% Noise
LSTM	MAE (kW)	0.0900	0.0902	0.0917	0.0951
	RMSE (kW)	0.1225	0.1228	0.1249	0.1302
	R2	0.8537	0.8531	0.8478	0.8346
TCN	MAE (kW)	0.0906	0.0913	0.0921	0.0976
	RMSE (kW)	0.1228	0.1236	0.1249	0.1318
	R2	0.8530	0.8512	0.8479	0.8305
MLP	MAE (kW)	0.0931	0.0934	0.0944	0.0987
	RMSE (kW)	0.1244	0.1249	0.1264	0.1326
	R2	0.8490	0.8479	0.8441	0.8285
XGBoost	MAE (kW)	0.1026	0.1087	0.111	0.1137
	RMSE (kW)	0.1328	0.1443	0.1477	0.1502
	R2	0.8280	0.7971	0.7874	0.7800

Table 8
Comparison of model predictive performance under Laplace noise.

Model	Metric	0%(Baseline)	5% Noise	10% Noise	20% Noise
LSTM	MAE (kW)	0.0900	0.0904	0.0912	0.0954
	RMSE (kW)	0.1225	0.1232	0.1246	0.1305
	R2	0.8537	0.8520	0.8487	0.8346
TCN	MAE (kW)	0.0906	0.0908	0.0919	0.0964
	RMSE (kW)	0.1228	0.1231	0.1246	0.1302
	R2	0.8530	0.8523	0.8487	0.8348
MLP	MAE (kW)	0.0931	0.0936	0.0947	0.0986
	RMSE (kW)	0.1244	0.1251	0.1269	0.1330
	R2	0.8490	0.8475	0.8429	0.8276
XGBoost	MAE (kW)	0.1026	0.1083	0.11	0.114
	RMSE (kW)	0.1328	0.1435	0.1454	0.1503
	R2	0.8280	0.7994	0.7940	0.7799

the risk that subsequent iterations overfit the noise.Although regularization is applied to mitigate overfitting, it may be insufficient in strong noise or complex heat load dataset, leading to poorer generalization and lower test accuracy. Overall, deep learning models exhibit superior robustness under noisy conditions. This is mainly attributed to their structural designs. LSTM gates effectively suppress high-frequency stochastic perturbations, while TCN's causal convolutions filter out such noise, allowing both models to maintain stability and predictive performance. In contrast, XGBoost is more sensitive to input noise, showing lower stability and generalization capability in complex noisy environments.

3.3. *Physical constraint post-processing mechanism*

Data-driven deep neural network models can effectively capture temporal dependencies. However, purely data-driven models still exhibit limitations in capturing highly variable and extreme operating conditions, and potentially leading to predictions that violate physical constraints. For instance, when a boiler is offline, its theoretical energy consumption should be zero, yet any nonzero prediction output represents invalid noise.To address this issue, introduces a physics-constrained (PC) post-processing mechanism based on boiler operating states. Specifically, a binary operating state variable, boiler_on, is defined according to the boiler start-stop status information. When boiler_on = 0, the corresponding prediction is constrained to zero, ensuring consistency between model outputs and physical boundary conditions. Unlike physics-informed neural networks (PINNs), which incorporate governing equations into the training loss function, the proposed PC mechanism acts as a lightweight post-processing constraint layer to enforce operational feasibility without requiring additional retraining or explicit thermodynamic modeling. This approach requires no additional training cost and can reduce systematic bias.

Table 9 summarizes the quantitative performance of the models with the PC mechanism. With PC applied, for LSTM, MAE and RMSE decrease by 28.2% and 15.3%, from 0.082 to 0.058 kW and 0.118 to 0.100 kW, respectively. For TCN, MAE and RMSE decrease by 35.1% and 19.4%, from 0.0864 to 0.0567 kW and 0.1196 to 0.0964 kW, respectively. Fig. 9 visually compares the predictive accuracy of all models under the PC mechanism, further demonstrating the performance improvement

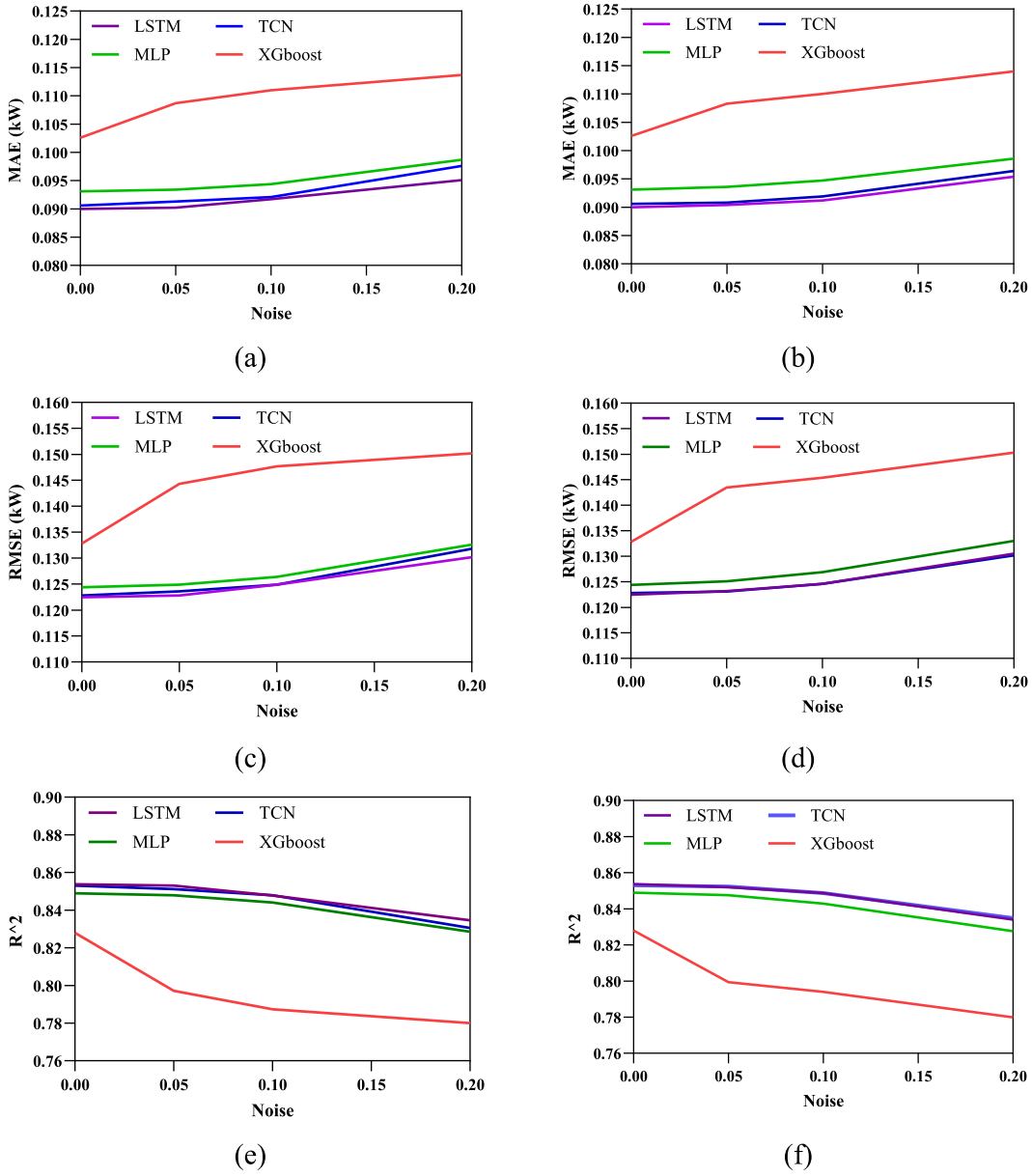


Fig. 8. Comparison of model prediction performance under different noise distributions and intensities. The left column (a, c, e) represents Gaussian noise scenarios, and the right column (b, d, f) represents Laplace noise scenarios.

Table 9
Models prediction performance with physical constraints.

Algorithms	MAE (kW)	RMSE (kW)	R^2
LSTM	0.058	0.1	0.9025
TCN	0.0567	0.0964	0.9093
MLP	0.0596	0.1007	0.9012
XGBoost	0.0618	0.1040	0.8946
LR	0.0644	0.1069	0.8885

achieved by incorporating physical constraints. These results indicate that the PC mechanism can not only mitigates the limitations of linear models in modeling complex nonlinear but also enhances predictive accuracy of deep temporal models.These results indicate that the PC mechanism not only mitigates the inherent limitations of linear models in capturing complex nonlinear dynamics but also enhances the predictive accuracy of deep temporal models. More importantly, it ensures that predictions remain physically consistent under extreme operating conditions, thereby increasing the practical applicability of the models for industrial scheduling and operational optimization.

3.4. Sensitivity analysis of LSTM-PC heat load prediction model

This study conducts a One-at-a-Time (OAT) sensitivity analysis to evaluate the influence of indoor temperature setpoint, outdoor temperature, and equipment load on a LSTM-PC prediction model. The sensitivity analysis is conducted by varying the deviation level of a single variable while keeping all other variables constant, thereby observing variations in model performance metrics. These three variables are selected because they represent the primary internal and external factors affecting building heat load. From a thermodynamic perspective, indoor temperature setpoint and outdoor temperature directly determine the temperature gradient driving heat transfer through the building envelope, while equipment load introduces internal heat gains that alter the building's energy balance.

Three deviation levels are defined in this study. At levels ±0.5, ±1,

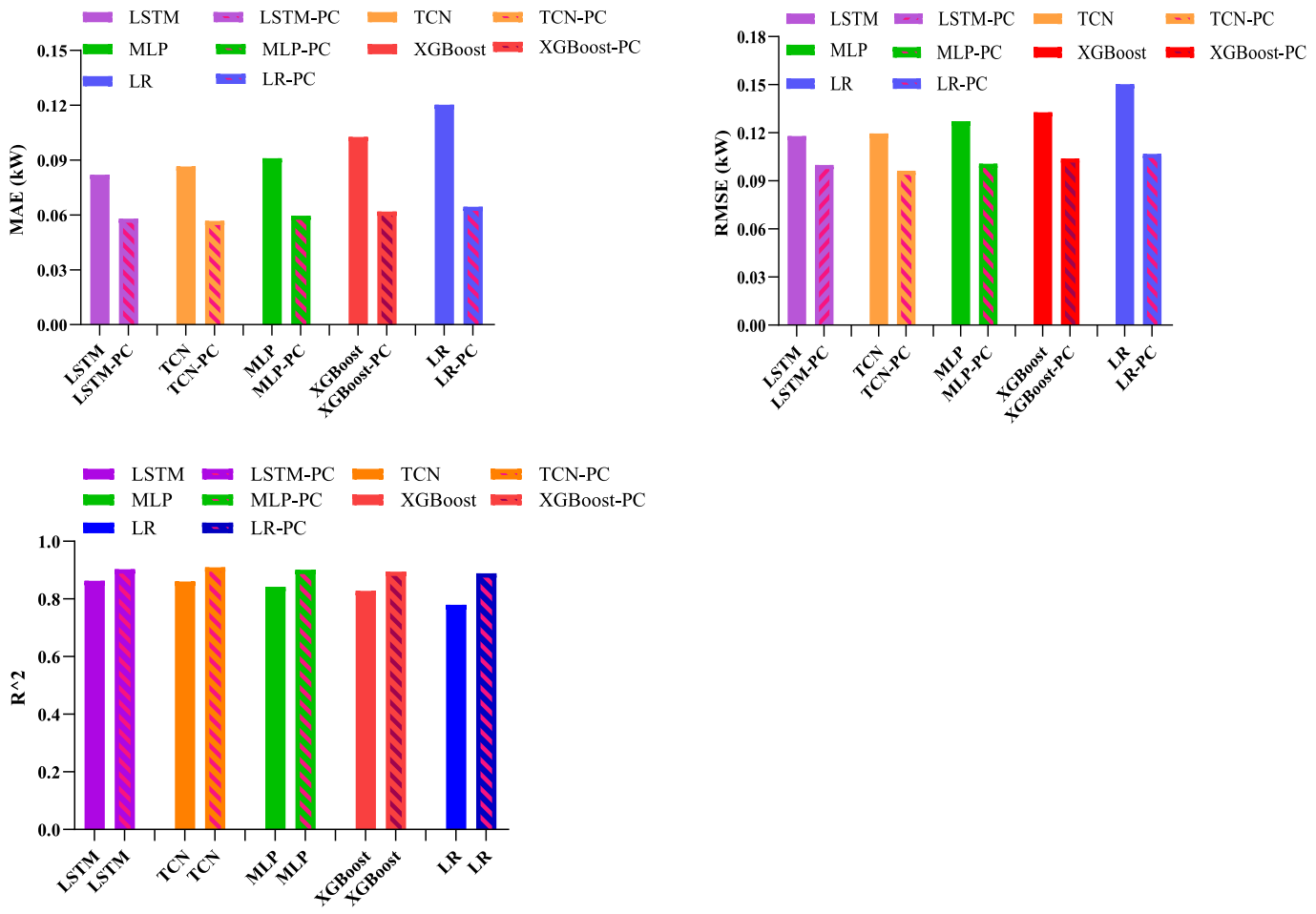


Fig. 9. Quantitative comparison performance of the prediction models using with and without the Physics-Constrained (PC) mechanism.

and ± 1.5, indoor setpoint and outdoor temperatures are perturbed by ±0.5 °C, ±1 °C, and ± 1.5 °C, respectively, while equipment load is varied by ±5%, ±10%, and ± 15%. As shown in Fig. 10(a)–(c), the results indicate clear differences in the model's 10 sensitivity to the three variables. Among them, the indoor temperature setpoint exhibits the highest sensitivity, followed by outdoor temperature with a moderate impact, while equipment load has the least influence. This suggests that the building heat load prediction model is primarily driven by temperature-related parameters, particularly the indoor temperature setpoint. From the MAE results in Fig. 10(a), the blue curve corresponding to indoor temperature setpoint shows a distinct U-shaped trend. The model achieves the lowest MAE at the baseline condition (Deviation Level = 0). As the indoor temperature setpoint deviates from the baseline, MAE increases significantly regardless of whether the deviation is positive or negative. At the highest deviation level (±1.5 °C), MAE increases from 0.058 kW to 0.0777 kW and 0.083 kW, respectively.

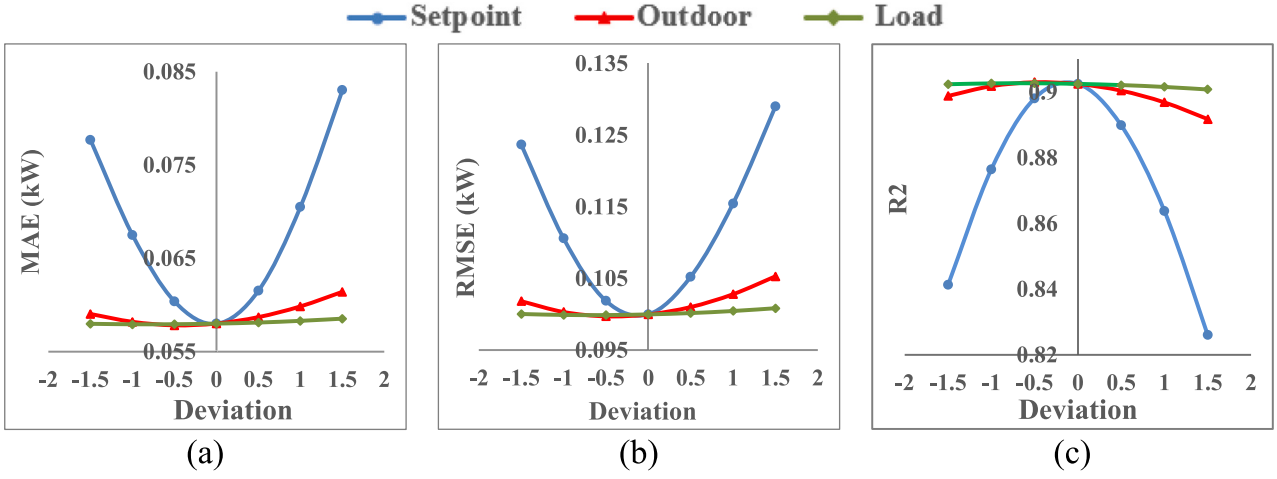


Fig. 10. OAT Sensitivity Analysis of LSTM-PC Heat Load Prediction Model: (a) MAE, (b) RMSE, and (c) R^2 under Different Deviation Levels of Indoor Temperature Setpoint, Outdoor Temperature, and Equipment Load.

Considering that the building heat load ranges from 0 to 2.045 kW, this level of error remains acceptable, indicating that the model maintains good predictive accuracy under perturbations. In contrast, outdoor temperature has a smaller impact on model performance than the indoor temperature setpoint. As the deviation in outdoor temperature increases, MAE show only slight upward trends, with relatively limited overall variation. This indicates that the model demonstrates a certain degree of robustness to external environmental disturbances. From a building thermal perspective, the thermal inertia of the building envelope can buffer short-term fluctuations in outdoor temperature, thereby weakening its direct influence on building heat load. The effect of equipment load on model performance is the weakest. Under different deviation conditions of equipment load, all performance metrics exhibit minor changes, indicating that the model is not sensitive to disturbances in internal equipment load. This may be because, compared with temperature-driven heat transfer processes, equipment load contributes a smaller proportion to the overall building heat load; therefore, its variations have limited impact on the total heat load prediction. The trends of RMSE in **Fig. 10 (b)** and R^2 in Fig. 10(c) are highly consistent with the MAE results, further confirming the sensitivity characteristics of the model with respect to different variables. Overall, the sensitivity analysis demonstrates that the proposed building heat load prediction model maintains relatively stable performance under disturbances in outdoor temperature and equipment load, indicating good robustness and generalization capability across different climatic conditions and operating scenarios. Meanwhile, the high sensitivity to indoor temperature setpoint is consistent with thermodynamic and building heat transfer principles, as the indoor and outdoor temperature difference is the primary driving force behind building heat exchange processes and HVAC energy demand.

4. Conclusion

This study focuses on a single apartment located in Bologna. Using the previous 24 h of data to predict the next hour heat load, an automated hyperparameter optimization framework based on Optuna was employed to systematically tune the models. A comparative analysis was conducted on five models, including LSTM, TCN, MLP, XGBoost, and LR, using heating data from March 6 to March 25, 2023. The main conclusions are as follows:

1) The comparative analysis indicates that sequence-based deep learning models, LSTM and TCN, achieve the best performance in heat load prediction, offering higher accuracy and robustness. They can effectively capture the temporal dependencies and thermal inertia of building heating systems, significantly outperforming traditional linear models. The MLP model ranks next, with lower computational cost and acceptable performance.Overall, LSTM and TCN are recommended for applications requiring high predictive accuracy.When computational resources are limited, MLP serves as an efficient alternative.
2) Noise robustness analysis indicates that as the intensity of Gaussian and Laplacian noise increases, the performance of all models degrades. However, LSTM and TCN consistently exhibit superior stability and resistance to noise. This is mainly attributed to their architectural designs: LSTM suppresses high-frequency noise propagation through its gating mechanism, while TCN smooths input disturbances via causal convolution, thereby maintaining stable temporal features. MLP shows some degree of robustness but with relatively higher errors, whereas XGBoost is more sensitive to noise and suffers significant performance degradation under high-noise conditions.
3) The post-processing mechanism based on operating states (PC) effectively eliminates unreasonable predictions during equipment shutdowns and significantly reduces prediction errors without additional training cost. Specifically, the MAE and RMSE of LSTM decrease by 28.2% and 15.3%, respectively, while those of TCN decrease by 35.1% and 19.4%. By integrating data-driven models with physical constraints, this approach not only ensures predictions comply with physical boundaries but also greatly enhances model reliability and applicability in real-world engineering scenarios.

5. Limitations and future work

Despite the proposed framework demonstrate high accuracy and strong robustness in heat load prediction, several limitations of this study should be acknowledged. First, the proposed framework was developed and evaluated using data from a single apartment located in Bologna, which does not validate the generalization capability of model across different building types and operating conditions. Second, the dataset used in this work is simulation-based rather than derived from real-world measurements, which may introduce discrepancies between simulated and practical deployment environments. In addition, the present study focuses mainly on hourly short-term forecasting, while the model performance over longer temporal horizons remains unverified. Furthermore, the transferability of the proposed framework across different climate zones, and diverse heating system configurations has not yet been systematically investigated. Future research should therefore include multi-building validation using real-world measured data, cross-climate and cross-system transferability analysis, and longer-term forecasting evaluation to further enhance the robustness and practical applicability of the proposed framework.

CRediT authorship contribution statement

Minghui Ma: Writing – original draft, Validation, Software, Investigation, Formal analysis, Data curation. **Paolo Valdiserri:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Vincenzo Ballerini:** Software, Methodology, Investigation. **Ruixin Li:** Writing – original draft, Software, Formal analysis. **Eugenia Rossi di Schio:** Writing – review & editing, Supervision, Methodology, Data curation, Conceptualization.

Declaration of competing interest

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

Appendix A. Estimation of Building Thermal Time Constant

The building thermal time constant is defined as:

$\tau = \frac{C}{H}$ (A1)where: C is the effective thermal capacity of the building (J K^{-1}); H is the total heat loss coefficient of the building (W K^{-1}). The total heat loss coefficient consists of transmission heat losses and ventilation heat losses:

$H = H_{tr} + H_{ve}$ (A2)

The transmission heat loss coefficient is calculated as:

$H_{tr} = \sum (U_i A_i)$ (A3)where: U_i is the thermal transmittance (U-value) of the i-th building envelope component (W m^{-2} K^{-1})); A_i is the corresponding component area (m^2).

The ventilation heat loss coefficient is expressed as:

$H_{ve} = \rho_{air} c_p \dot{V}$ (A4)where: ρ_{air} is the air density (kg m^{-3}); c_p is the specific heat capacity of air at constant pressure (J kg^{-1} K^{-1}); $\dot{V}$is the ventilation volumetric flow rate (m^3 s^{-1}).

The volumetric flow rate can be further expressed as a function of the air change rate and room volume:

$\dot{V} = \frac{nV}{3600}$ (A5)where: n is the air change rate (h^{-1}); V is the room volume (m^3).

By substituting standard air properties ($\rho_{air} \approx 1.20$ kg m^{-3}, $c_p \approx 1005$ J kg^{-1} K^{-1}), the ventilation heat loss coefficient can be simplified as:

$H_{ve} = 0.33nV$ (A6)

The effective thermal capacity of the building is calculated as:

$C = \sum_j \rho_j c_j A_j d_j$ (A7)where: d_j is the thickness of the j-th aterial layer (m); ρ_j is the material density (kg m^{-3}); c_j is the specific heat capacity of the material (J kg^{-1} K^{-1})); A_j is the material area (m^2). The effective thermal capacity depends on the thermally activated layer near the internal surfaces. According to ISO 52016-1:2017, an effective thickness of 10 cm was adopted in this study.

Data availability

The data and code that support the findings of this study are available from the corresponding author upon reasonable request.

References

[1] International Energy Agency, World Energy Outlook 2024, IEA, 2024.
[2] International Energy Agency, CO2 Emissions in 2023, IEA, 2024.
[3] I.L. Niza, I.M.D. Luz, A.M. Bueno, E.E. Broday, Thermal comfort and energy efficiency: challenges, barriers, and step towards sustainability, Smart Cities 5 (4) (2022) 1721–1741.
[4] L. Zhang, J. Wen, Y. Li, J. Chen, Y. Ye, Y. Fu, W. Livingood, A review of machine learning in building load prediction, Appl. Energy 285 (2021) 116452.
[5] R. Olu-Ajayi, H. Alaka, I. Sulaimon, F. Sunmola, S. Ajayi, Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques, J. Build. Eng. 45 (2022) 103406.
[6] J. Song, L. Zhang, G. Xue, Y. Ma, S. Gao, Q. Jiang, Predicting hourly heating load in a district heating system based on a hybrid CNN-LSTM model, Energ. Buildings 243 (2021) 110998.
[7] H.-X. Zhao, F. Magoulès, A review on the prediction of building energy consumption, Renew. Sust. Energ. Rev. 16 (6) (2012) 3586–3592.
[8] J. Yu, W.-S. Chang, Y. Dong, Building energy prediction models and related uncertainties: a review, Buildings 12 (8) (2022) 1284.
[9] K. Amasyali, N.M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renew. Sust. Energ. Rev. 81 (2018) 1192–1205.
[10] Q. Zhou, S. Wang, X. Xu, F. Xiao, A grey-box model of next-day building thermal load prediction for energy-efficient control, Int. J. Energy Res. 32 (15) (2008) 1418–1431.
[11] J. Sun, M. Gong, Y. Zhao, C. Han, L. Jing, P. Yang, A hybrid deep reinforcement learning ensemble optimization model for heat load energy-saving prediction, J. Build. Eng. 58 (2022) 105031.
[12] W. Jiang, P. Wang, X. Ma, Y. Liu, Development of a grey-box heat load prediction model by subspace identification method for heating building, Build. Environ. 280 (2025) 113119.
[13] Y. Guo, J. Wang, H. Chen, G. Li, J. Liu, C. Xu, R. Huang, Y. Huang, Machine learning-based thermal response time ahead energy demand prediction for building heating systems, Appl. Energy 221 (2018) 16–27.
[14] P.J.C. Vogler-Finck, Forecast and Control of Heating Loads in Receding Horizon, Aalborg Universitet, 2018.
[15] K. Yun, R. Luck, P.J. Mago, H. Cho, Building hourly thermal load prediction using an indexed ARX model, Energ. Buildings 54 (2012) 225–233.
[16] G. Ciulla, A. D'Amico, Building energy performance forecasting: a multiple linear regression approach, Appl. Energy 253 (2019) 113500.
[17] T. Ahmad, H. Chen, R. Huang, G. Yabin, J. Wang, J. Shair, H.M. Azeem Akram, S. A. Hassnain Mohsan, M. Kazim, Supervised based machine learning models for short, medium and long-term energy prediction in distinct building environment, Energy 158 (2018) 17–32.
[18] M. Ma, O. Pektezel, V. Ballerini, P. Valdiserri, E. Rossi di Schio, Performance predictions of solar-assisted heat pumps: methodological approach and comparison between various artificial intelligence methods, Energies 17 (22) (2024) 5607.
[19] S. Seyedzadeh, F.P. Rahimian, I. Glesk, M. Roper, Machine learning for estimation of building energy consumption and performance: a review, Visualization Eng. 6 (1) (2018) 5.
[20] J. Guo, S. Yun, Y. Meng, N. He, D. Ye, Z. Zhao, L. Jia, L. Yang, Prediction of heating and cooling loads based on light gradient boosting machine algorithms, Build. Environ. 236 (2023) 110252.
[21] N.-T. Ngo, A.-D. Pham, T.T.H. Truong, N.-S. Truong, N.-T. Huynh, T.M. Pham, An ensemble machine learning model for enhancing the prediction accuracy of energy consumption in buildings, Arab. J. Sci. Eng. 47 (4) (2022) 4105–4117.
[22] M.W. Ahmad, M. Mourshed, Y. Rezgui, Trees vs neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption, Energ. Buildings 147 (2017) 77–89.
[23] S. Chatterjee, S. Bayer, A. Maier, Prediction of Household-level Heat-Consumption using PSO Enhanced SVR Model, arXiv preprint arXiv:2112.01908, 2021.
[24] C. Chang, G. Ma, J. Zhang, J. Tao, Investigation on the CNN-LSTM-MHA-based model for the heating energy consumption prediction of residential buildings considering active and passive factors, Energy 333 (2025) 137508.
[25] Y. Wang, C. Zhan, G. Li, D. Zhang, X. Han, Physics-guided LSTM model for heat load prediction of buildings, Energ. Buildings 294 (2023) 113169.
[26] Y. Wang, C. Zhan, G. Li, S. Ren, Comparison of algorithms for heat load prediction of buildings, Energy 297 (2024) 131318.
[27] I. Karijadi, S.-Y. Chou, A hybrid RF-LSTM based on CEEMDAN for improving the accuracy of building energy consumption prediction, Energ. Buildings 259 (2022) 111908.
[28] J. Zhang, Y. Huang, H. Cheng, H. Chen, L. Xing, Y. He, Ensemble learning-based approach for residential building heating energy prediction and optimization, J. Build. Eng. 67 (2023) 106051.
[29] A. Zhao, L. Mi, X. Xue, J. Xi, Y. Jiao, Heating load prediction of residential district using hybrid model based on CNN, Energ. Buildings 266 (2022) 112122.
[30] J. Jang, J. Han, S.-B. Leigh, Prediction of heating energy consumption with operation pattern variables for non-residential buildings using LSTM networks, Energ. Buildings 255 (2022) 111647.
[31] S. Ardabili, L. Abdolalizadeh, C. Mako, B. Torok, A. Mosavi, Systematic review of deep learning and machine learning for building energy, Front. Energy Res. 10 (2022) 786027.
[32] Y. Ding, Q. Zhang, T. Yuan, F. Yang, Effect of input variables on cooling load prediction accuracy of an office building, Appl. Therm. Eng. 128 (2018) 225–234.
[33] V. Ballerini, E.P.B. de Volo, B. Pulvirenti, E.R. di Schio, P. Valdiserri, P. Guidorzi, Influence of different heating systems on thermal comfort perception: A dynamic and CFD analysis, in: Journal of Physics: Conference Series vol. 2685, IOP Publishing, 2024, 012021.
[34] S.A. Klein, TRNSYS 18: A Transient System Simulation Program, Solar Energy Laboratory, University of Wisconsin, Madison, USA, 2017.
[35] H.M. Dott, J. Ruschenburg, F. Ochs, J. Bony, The Reference Framework for System Simulations of the IEA SHC Task 44 / HPP Annex 38, International Energy Agency (IEA) Solar Heating and Cooling Programme (SHC), 2013. Report C1 Part B.