Journal of Engineering and Applied Science

RESEARCH Open Access

Research on cooling load estimation through optimal hybrid models based on Naive Bayes

Ying Xu[1*]

*Correspondence:
jiayouxy131425@163.com

[1] Basic Department, The Tourism College of Changchun University, Changchun 130000, Jilin, China

Abstract

Cooling load estimation is crucial for energy conservation in cooling systems, with applications like advanced air-conditioning control and chiller optimization. Traditional methods include energy simulation and regression analysis, but artificial intelligence outperforms them. Artificial intelligence models autonomously capture complex patterns, adapt, and scale with more data. They excel at predicting cooling loads influenced by various factors, like weather, building materials, and occupancy, leading to dynamic, responsive predictions and energy optimization. Traditional methods simplify real-world complexities, highlighting artificial intelligence's role in precise cooling load forecasting for energy-efficient building management. This study evaluates Naive Bayes-based models for estimating building cooling load consumption. These models encompass a single model, one optimized with the Mountain Gazelle Optimizer and another optimized with the horse herd optimization algorithm. The training dataset consists of 70% of the data, which incorporates eight input variables related to the geometric and glazing characteristics of the buildings. Following the validation of 15% of the dataset, the performance of the remaining 15% is tested. Based on analysis through evaluation metrics, among the three candidate models, Naive Bayes optimized with the Mountain Gazelle Optimizer (NBMG) demonstrates remarkable accuracy and stability, reducing prediction errors by an average of 18% and 31% compared to the other two models (NB and NBHH) and achieving a maximum R^2 value of 0.983 for cooling load prediction.

Keywords: Cooling load estimation, Prediction models, Building energy consumption, Naive Bayes, Metaheuristic optimization algorithms

Introduction

In the contemporary era, the escalating demand for energy, primarily from residential and commercial sectors, poses challenges in efficiently managing industries like transportation and construction while striving to conserve energy [1, 2]. Recent studies emphasize the substantial contribution of a growing population to energy consumption in residential buildings [3, 4]. Efficiently managing a building's energy consumption requires a thorough understanding of its performance, starting with the identification of energy sources and usage patterns. Key energy resources in buildings include district

heating supply, electricity, and natural gas, with applications such as HV heating, ventilation, and air-conditioning (HVAC) systems, lighting, elevators, hot water, and kitchen equipment consuming this energy [1]. Among these, HVAC systems, important for residential infrastructure, significantly impact cooling load (CL) and heating load (HL), constituting around 40% of energy consumption in office buildings [5, 6]. Improving energy efficiency in urban residential buildings and employing dynamic load prediction in construction management are crucial measures to enhance HVAC system performance and conserve energy [7]. Forecasting dynamic air-conditioning loads is essential for HVAC system design, enabling adjustments to initiation times, curbing peak demand, optimizing costs, and improving energy utilization in cooling storage systems [8]. Accurately predicting building cooling loads is challenging due to various influencing factors, including optical and thermal characteristics and meteorological data [9–11].

Achieving sustainability in thermal management relies on efficiently separating latent and sensible loads in the cooling process. An effective strategy involves integrating an indirect evaporative cooler (IEC) with a dehumidification system, providing both enhanced cooling efficiency and a sustainable solution to rising energy demands. The improved IEC, featuring three significant modifications, becomes a cornerstone in this approach, pushing the coefficient of performance (COP) for cooling to an impressive 78. The dehumidification component, operating at a COP of approximately 4–5, complements the cooling-only COP, resulting in an overall COP of 7–8 [12].

Efforts to create energy-efficient buildings and enhance energy conservation are necessary in managing energy demand and resources. A primary strategy involves early predictions of HL and CL in residential structures. Accurate forecasting requires data on building specifications and local weather conditions [13]. Climatic elements such as temperature, wind speed, solar radiation, atmospheric pressure, and humidity significantly influence the prediction of building cooling and heating loads. Factors like relative compactness, roof dimensions, wall and glazing areas, roof height, and overall surface area should be considered when assessing a building's load [14]. Building energy simulation tools play a crucial role in designing energy-efficient buildings, allowing for performance maximization and comparisons between buildings. Simulation outcomes have demonstrated high accuracy in replicating real-world measurements [15]. Although time-intensive and requiring proficient users, simulation software effectively assesses the influence of building design factors. In some cases, contemporary techniques like statistical analysis, artificial neural networks, and machine learning are adopted to predict cooling and heating loads and analyze the impact of different parameters [16].

HVAC system optimization involves three main categories: simulation, regression analysis, and artificial intelligence (AI). Simulation tools like DOE-2 [17], ESP-r [9], TRNSYS [10], and EnergyPlus [11, 18] are utilized for cooling load estimation when comprehensive building data is available. However, challenges arise in accurately measuring various parameters, and simplifying building models demands significant time and resources [19]. Simulation software is limited to real-time applications like online prediction or optimal operational control [20]. Regression analysis, known for its ease of use and computational efficiency, is preferred for diverse building types [21], employing both linear and nonlinear techniques [22, 23]. Additionally, research emphasizes the efficacy of ML and AI in building energy forecasting, favoring nonlinear approaches

[24, 25]. Building cooling load prediction commonly involves key factors such as outdoor temperature, relative humidity, solar irradiation, and indoor occupancy schedules [26, 27]. Feature extraction methods, including engineering, statistical, and structural approaches, help condense raw data into informative formats, addressing the complexity introduced by historical data [21].

Numerous data mining methods have been applied to predict residential building energy requirements, including principal component analysis (PCA) [28], extreme learning machine (ELM) [29, 30], support vector machines (SVM) [31–33], k-means [34], deep learning [32, 33, 35–37], decision trees (DT) [38], various regression approaches, artificial neural networks [16, 39, 40], and hybrid models [41–44]. Researchers have employed diverse methodologies to forecast heating and cooling loads and energy demand in various building contexts. For instance, one study [45] predicted building heating load using the MLP method with meteorological data, while another simultaneously [46] predicted both cooling and heating loads with meteorological and date data inputs. Another study [16] examined a building's energy performance using machine learning techniques, including general linear regression, artificial neural networks, decision trees, support vector regression (SVR), and ensemble inference models for cooling and heating load forecasting. Structural and interior design factors' impact on cooling loads was explored through diverse regression models [47], and HVAC system energy demand was estimated from cooling and heating load requirements using different regression models. Commercial buildings' cooling load and electric demand were forecasted for short-term and ultrashort-term management [48], enhancing energy efficiency through a hybrid SVR approach. Additionally, the SVR method was applied [49] to project cooling loads in a large coastal office building in China, introducing a novel vector-based SVR model for increased robustness and forecasting precision [50].

Naive Bayes is a fundamental probabilistic machine learning algorithm widely employed in various fields, including natural language processing, spam filtering, and classification tasks. It is rooted in Bayes' theorem and assumes conditional independence between features, which is where the "naive" in its name originates. This simplifying assumption enables Naive Bayes to efficiently estimate the probability of a data point belonging to a particular class. Despite its simplicity, Naive Bayes often exhibits impressive classification performance, especially when dealing with high-dimensional and large datasets. To date, there is no article to use Naïve Bayes as the prediction model in the case of CL of the buildings. In this study, Naïve Bayes single model prediction performance is compared with two optimized counterparts (optimized with Mountain Gazelle Optimizer (MGO) and the horse herd optimization algorithm (HHO)). The following sections present an academic description of the model and selected optimizers and a comparative analysis between developed models.

Methods

Data collection

The main goal of this study is to forecast the cooling load (CL) in buildings. This is achieved by using experimental data extracted from energy consumption patterns documented in previous studies [51, 52]. Table 1 reports the statistical properties (minimum, maximum, average, and standard deviation) of the variables included in the training of

Table 1 The statistic properties of the input variable of NB [51, 52]

Variables		Indicators			
	Category	Min	Max	Avg	St. dev
Relative compactness	Input	0.62	0.98	0.764	0.106
Surface area (m^2)	Input	514.5	808.5	671.7	88.09
Wall area (m^2)	Input	245	416.5	318.5	43.63
Roof area (m^2)	Input	110.25	220.5	176.6	45.17
Overall height (m)	Input	3.5	7	5.25	1.751
Orientation	Input	2	5	3.5	1.119
Glazing area (%)	Input	0	0.4	0.234	0.133
Glazing area distribution	Input	0	5	2.813	1.551
Cooling (KW)	Output	10.9	48.03	24.59	9.513

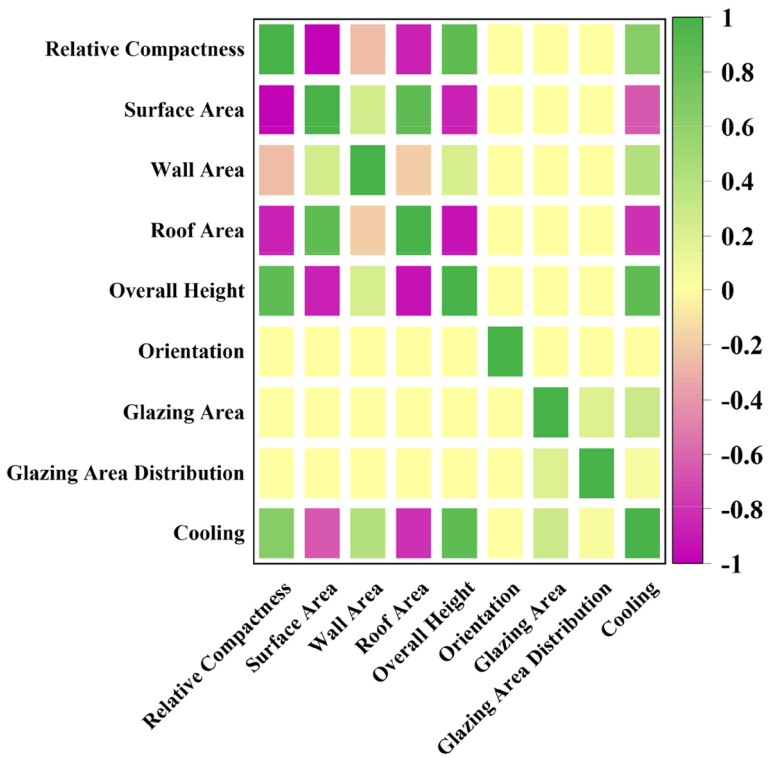

Fig. 1 The correlation between input and output parameters

the developed prediction models and the output. Input parameters include relative compactness (indicating the building's surface area-to-volume ratio), surface area, roof area, wall area, orientation, overall height, glazing area (encompassing glazing, frame, and sash components), and the distribution of glazing area, and cooling load is the expected output variable.

Figure 1 visually represents the correlation among the variables examined in this study. The analysis depicted in the figure reveals compelling insights. Specifically, it becomes apparent that the overall height and relative compactness exhibit the most substantial positive impact on the cooling load. In contrast, roof area and surface area emerge as

variables with the most pronounced negative influence on the cooling load. This graphical representation not only highlights the interrelationships between the variables but also emphasizes the varying degrees of impact each variable has on the cooling load.

Overview of machine learning methods and optimizers

Naive Bayes (NB)

The Naive Bayes (NB) classifier stands as a robust probabilistic model founded on Bayes' theorem, which simplifies modeling by assuming independence among input variables. Its potential for substantial improvements in prediction accuracy becomes evident when combined with kernel density approximations, as highlighted in [53, 54].

The NB is a sophisticated system that smoothly integrates the Naive Bayes probability model into its decision-making process. This classifier relies on the maximum a posteriori (MAP) decision rule, a well-established method for identifying the most probable hypothesis from a given set of options. Additionally, there is a closely related classifier called the Bayes classifier. This robust algorithm is responsible for assigning class labels $y = C_k$, where k can range from 1 to K. This involves a detailed evaluation of various factors and variables, leading to the categorization of data points into predefined classes.

$$y = argmaxp(C_k)\prod_{i=1}^{n} p((x_i|C_k)) \tag{1}$$

In the provided equation, the variable y represents the predicted class label assigned by the Naive Bayes classifier. The term C_k denotes a specific class, where k ranges from 1 to K, indicating the total number of classes. The variable n represents the total number of input features or variables, and x_i refers to the $i - th$ input feature or variable. The term $p(C_k)$ represents the prior probability of class C_k, while $p(x_i \mid C_k)$ denotes the conditional probability of observing x_i given the class C_k.

Mountain gazelle optimizer (MGO)

The MGO algorithm is inspired by the behavior of mountain gazelles, which are grouped into bachelor herds, maternity herds, and solitary, territorial males. It aims to find optimal solutions by designating adult male gazelles in herd territories as global optima. Mathematically defined, the algorithm balances exploitation and exploration, gradually moving toward optimal solutions using four specified exploration mechanisms [55].

Territorial solitary males Mature mountain gazelles establish solitary territories, vigorously defending them from other males seeking access to females. Equation (2) models these territories.

$$TSM = male_{gzl} - |(ri_1 \times YH - ri_2 \times X(t)) \times F| \times Cof_r \tag{2}$$

Equation (2) describes $male_{gzl}$ as the adult man is the most effective overall solution, as seen by the position vector. The variables ri_2 and ri_1 are random integers that can take on a value of either 1 or 2 [55]. YH denoted the coefficient vector of utilizing Eq. (3), and one can determine the young male herd. Similarly, F is computed using Eq. (4). In each iteration, the coefficient vector Cof_r, selected at random, undergoes

updates and is employed to augment the search capability. This coefficient vector is specified using Eq. (3).

$$YH = X_{ra} \times \lfloor r_1 \rfloor + M_{pr} \times \lceil r_2 \rceil, ra = \left\{ \lceil \frac{N}{3} \rceil \ldots N \right\} \tag{3}$$

Here, X_{ra} denotes a random solution (young 1 male) within the range of ra. M_{pr} refers to the average number of search agents, which is equal to $\lceil \frac{N}{3} \rceil$, and N is the total number of gazelles, while r_1 and r_2 are random values in [0, 1].

$$F = N_1(D) \times \exp\left(2 - Iter \times \left(\frac{2}{MaxIter}\right)\right) \tag{4}$$

Equation (4) incorporates multiple variables associated with the problem's dimensions. A randomly generated number following a standard distribution denoted as N_1 and exp is the equation that employs the exponential function. $Iter$ shows the ongoing iteration number in the process, and $MaxIter$ signifies the total count of iterations.

$$Cof_i = \begin{cases} (x+1) + r_3, \\ x \times N_2(D), \\ r_4(D), \\ N_3(D) \times N_4(D)^2 \times \cos((r_4 \times 2) \times N_3(D)), \end{cases} \tag{5}$$

$$x = -1 + Iter \times \left(\frac{-1}{MaxIter}\right) \tag{6}$$

Additionally, r_3, r_4, and $rand$ are random numbers from 0 to 1 [55]. N_2, N_3, and N_4 denote random numbers drawn from a typical distribution, and it is related to the dimensions of the problem. $Iter$ indicates the current iteration number, while $MaxIter$ is the number of iterations to be performed.

Maternity herds Maternity herds hold a crucial position within the mountain gazelles' life cycle since they are principally responsible for producing strong male gazelles. Furthermore, male gazelles may actively participate in the delivery process of the offspring and confront the presence of younger males attempting to mate with females. This behavioral interplay is expressed mathematically in Eq. (7).

$$MH = \left(YH + Cof_{1,r}\right) + (ri_3 \times male_{gzl} - ri_4 \times X_{rand}) \times Cof_{1,r} \tag{7}$$

Here, YH signifies the young men's impact factor vector, which is determined by using Eq. (3). $Cof_{2,r}$ and $Cof_{3,r}$ random vectors for the coefficients are determined independently using Eq. (5). ri_3 and ri_4 are random integers that can take on a value of either 1 or 2. $male_{gzl}$ denoted the best global solution (adult male) in the current iteration. Ultimately, X_{rand} corresponds to the location vector of a gazelle chosen at random from the entire herd.

Bachelor male herds Male gazelles create territories after they reach adulthood and engage in mating pursuit, a period marked by intense competition between young and

adult males for territory control and access to females, as mathematically captured in Eq. (8).

$$YMH = (X(t) - D) + (ri_5 \times male_{gazelle} - ri_6 \times YH) \times Cof_r \tag{8}$$

$$D = (|X(t)| + |male_{gzl}|) \times (2 \times r_6 - 1) \tag{9}$$

where $X(t)$ indicates the gazelle's current iteration's location vector. The variables ri_5 and ri_6 are random integers that can take a value of either 1 or 2. The ideal answer designates the male gazelle's location vector as $male_{gzl}$. r_6 is also a random number from 0 to 1.

Migration to search for food Equation (10), which describes how mountain gazelles forage for food, takes into account their extraordinary sprinting and leaping speed.

$$MSF = (ul - ll) \times r_7 + ll \tag{10}$$

where ul and ll represent the lower and upper limits of the problem, respectively. Furthermore, r_7 is a random integer in [0, 1], and it is selected randomly.
The pseudo-code of MGO is available as follows:

%*MGO* setting
Inputs: The population size *N* and maximum number of iterations *I*
Outputs: Gazelle's location and fitness potential

% initialization
Create a random population using $X_i (i = 1, 2, ..., N)$
Calculate the gazelle's fitness level
While (the stopping condition is not met) do
For (each gazelle (X_i)) do

% Alone male realm
Calculate TSM using Eq. (2)

% Mother and child herd
Calculate *MH* using Eq. (7)

% Young male herd
Calculate *YMH* using Eq. (8)

% Migration to search for food
Calculate *MSF* using Eq. (10)
Calculate the fitness values of *TSM*, *MH*, *YMH*, and *MSF* and then add them to the habitat
End for
Sort the entire population in ascending order
Update $best_{\text{Gazelle}}$
Save the *N* best gazelles in the max number of population
end, while
Return $X_{\text{BestGazelle}}$*best Fitness*

Horse herd optimization algorithm (HOA) The HOA is based on how horses behave in the wild [56]. This information is based on six specific behaviors: grazing, hierarchy, imitation, sociability, roaming, and defense mechanisms. These actions are the foundation of HOA, directing the movement of horses in each cycle, as detailed in Eq. (11):

$$X_m^{Iter,A} = \overrightarrow{V}_m^{Iter,A} + X_m^{(Iter-1),A}, A(Age) = \alpha, \beta, \gamma, \delta \tag{11}$$

where $X_m^{Iter,A}$ denotes the position of the $m - th$ horse, A represents the age range, and *Iter* is the current iteration. A also reflects the horse's age range, while $\overrightarrow{V}_m^{Iter,A}$ indicates the velocity vector of the horse. Horses typically live between 25 and 30 years, exhibiting

various behaviors throughout their lifespan. These behaviors are categorized into δ (0–5 years), γ (5–10 years), and α (older than 15 years) groups. An extensive response matrix determines how old horses are sorted by how well they perform. The top 10% form group α, the next 20% belong to group β, and the remaining 30% and 40% are categorized as groups γ and δ, respectively. Motion vectors corresponding to equines of varying age groups and computational cycles within the algorithm are established following these behavioral patterns.

$$\begin{aligned}
\overrightarrow{V}_m^{Iter,\alpha} &= \overrightarrow{G}_m^{Iter,\alpha} + \overrightarrow{D}_m^{Iter,\alpha} \\
\overrightarrow{V}_m^{Iter,\beta} &= \overrightarrow{G}_m^{Iter,\beta} + \overrightarrow{H}_m^{Iter,\beta} + \overrightarrow{S}_m^{Iter,\beta} + \overrightarrow{D}_m^{Iter,\beta} \\
\overrightarrow{V}_m^{Iter,\gamma} &= \overrightarrow{G}_m^{Iter,\gamma} + \overrightarrow{H}_m^{Iter,\gamma} + \overrightarrow{S}_m^{Iter,\gamma} + \overrightarrow{I}_m^{Iter,\gamma} + \overrightarrow{D}_m^{Iter,\gamma} + \overrightarrow{R}_m^{Iter,\gamma} \\
\overrightarrow{V}_m^{Iter,\delta} &= \overrightarrow{G}_m^{Iter,\delta} + \overrightarrow{I}_m^{Iter,\delta} + \overrightarrow{R}_m^{Iter,\delta}
\end{aligned} \tag{12}$$

To elucidate the derivation of the global matrix, Eqs. (13) and (14) are utilized, and a relationship between positions (X) and their respective cost values ($C(X)$) is established.

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,d} \end{bmatrix}, C(X) = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix} \tag{13}$$

$$Global\ Matrix = [XC(X)] = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} & c_1 \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} & c_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,d} & c_m \end{bmatrix} \tag{14}$$

Here, m indicates the count of horses, and d is the dimensions of the problem. After that, the global matrix is arranged according to the final column, which signifies costs. The horse's age is recorded in this column. The velocity of horses under 5 years age range is as follows:

$$\begin{aligned}
\overrightarrow{V}_m^{Iter,\delta} = &\left[g_m^{(Iter-1),\delta}\omega_g\left(\breve{u} + P\breve{l}\right)\left[X_m^{(Iter-1)}\right]\right] + [i_m^{(Iter-1),\delta}\omega_i[(\frac{1}{P^N}\sum_{j=1}^{P^N}\widehat{X}_j^{Iter-1}) \\
&- X^{Iter-1}]]] + [r_m^{(Iter-1),\delta}\omega_r P X^{Iter-1}
\end{aligned} \tag{15}$$

The velocity of horses between 5 and 10 years age range:

$$\begin{aligned}
\overrightarrow{V}_m^{Iter,\gamma} = &\left[g_m^{(Iter-1),\gamma}\omega_g\left(\breve{u} + P\breve{l}\right)\left[X_m^{(Iter-1)}\right]\right] + [h_m^{(Iter-1),\gamma}\omega_h\left[X_*^{(Iter-1)} - X_m^{(Iter-1)}\right]] \\
&+ [S_m^{(Iter-1),\gamma}\omega_S[(\frac{1}{N}\sum_{j=1}^{N}X_j^{Iter-1}) - X^{Iter-1}]] + [i_m^{(Iter-1),\gamma}\omega_i[(\frac{1}{P^N}\sum_{j=1}^{P^N}\widehat{X}_j^{Iter-1}) - X^{Iter-1}]] \\
&- [d_m^{(Iter-1),\gamma}\omega_d[(\frac{1}{q^N}\sum_{j=1}^{q^N}\widehat{X}_j^{Iter-1}) - X^{Iter-1}]] + [r_m^{(Iter-1),AGE}\omega_r P X^{Iter-1}]
\end{aligned} \tag{16}$$

The velocity of horses between 10 and 15 years age range:

$$\begin{aligned}\overrightarrow{V}_m^{Iter,\beta} = &\left[g_m^{(Iter-1),\beta}\omega_g\left(\breve{u}+P\breve{l}\right)\left[X_m^{(Iter-1)}\right]\right]+[h_m^{(Iter-1),\beta}\omega_h\left[X_*^{(Iter-1)}-X_m^{(Iter-1)}\right]]\\&+[S_m^{(Iter-1),\beta}\omega_S[(\frac{1}{N}\sum_{j=1}^{N}X_j^{Iter-1})-X^{Iter-1}]]-[d_m^{(Iter-1),\beta}\omega_d[(\frac{1}{q^N}\sum_{j=1}^{q^N}\breve{X}_j^{Iter-1})-X^{Iter-1}]]\end{aligned} \tag{17}$$

Horses that are 15 years or older exhibit the following velocity:

$$\overrightarrow{V}_m^{Iter,\alpha} = \left[g_m^{(Iter-1),\alpha}\omega_g\left(\breve{u}+P\breve{l}\right)\left[X_m^{(Iter-1)}\right]\right]-[d_m^{(Iter-1),\alpha}\omega_d[(\frac{1}{q^N}\sum_{j=1}^{q^N}\breve{X}_j^{Iter-1})-X^{Iter-1}]] \tag{18}$$

Results and discussion

Hyperparameter results

External configurations referred to as hyperparameters—such as alpha and binarize—are important in shaping a model's behavior. Distinguished from parameters, these hyperparameters are predetermined and not acquired through the learning process of the data. The optimization of model performance significantly relies on the fine-tuning of hyperparameters, a nuanced process that demands both experimentation and the strategic application of optimization techniques. Table 2 outlines the hyperparameter values for the NBMG and NBHH models. By providing intricate insights into the intricacies of hyperparameter configurations, it becomes an indispensable tool for comprehending and, crucially, reproducing model setups. This exposition not only elevates the technical aspects of the research but also contributes to the broader scholarly discourse in the field of machine learning.

Prediction performance analysis

The assessment of the predictive effectiveness of the constructed models involved the utilization of five distinct metrics, which relied on actual observed values (T_i) and corresponding predicted values (P_i). Here, the symbols $\overline{T}$ and $\overline{P}$ denote the mean of all the outcomes subjected to testing and predicting. In contrast, n signifies the total count of samples encompassed within the analyzed dataset. A description of these metrics is presented as follows:

(1) The coefficient of determination (R^2) numerically represents the portion of the variability in the dependent variable that can be anticipated through the independent variables integrated into the model.

Table 2 The results of hyperparameters for NB

Models	Hyperparameter	
	Alpha	Binarize
NBMG	7	5.61
NBHH	6.52953	1.317853

$$R^2 = \left(\frac{\sum_{i=1}^{n}(T_i - \overline{T})(P_i - \overline{P})}{\sqrt{\left[\sum_{i=1}^{n}(T_i - \overline{P})^2\right]\left[\sum_{i=1}^{n}(P_i - \overline{P})^2\right]}} \right)^2 \tag{19}$$

(2) Root-mean-square error (RMSE) denotes the square root of the squared disparities' mean between the projected and observed values. This quantifies the typical magnitude of the discrepancies the model introduces when forecasting the target variable.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - T_i)^2}{n}} \tag{20}$$

(3) Mean squared error (MSE) calculates the average of the squared differences between predicted and actual values, measuring how well a model's predictions match the actual data. Lower MSE values indicate better predictive accuracy and a closer fit to the observed data.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(P_i - T_i)^2 \tag{21}$$

(4) Nash–Sutcliffe efficiency (NSE) assesses how well a model's predictions match observed values, considering the variability of the observed data. Higher NSE values indicate better model performance, with 1 indicating a perfect match.

$$NSE = 1 - \frac{\sum_{i=1}^{n}(P_i - T_i)^2}{\sum_{i=1}^{n}(T_i - \overline{T})^2} \tag{22}$$

(5) MDAPE (mean directional absolute percentage error) expresses the average percentage difference between the predicted and actual values, considering the direction of the errors (underestimation or overestimation).

$$RAE = \frac{\sum_{i=1}^{n}|P_i - T_i|}{\sum_{i=1}^{n}|T_i - \overline{T}|} \tag{23}$$

The following discussion comprehensively analyzes the model's performance in predicting CL based on Table 3:

- *NB (single model)*: A minimum R^2 value of 0.963 is reported for this model. High error values of ($RMSE = 2.147$, $MSE = 4.610$, and $MDAPE = 7.482$) indicated low accuracy of this traditional model, especially in the testing phase. Low NSE values of 0.966, 0.958, and 0.949 in the training, validation, and testing phases confirm the high variability of estimated data.
- *NBMG (NB + MGO)*: High R^2 values of 0.986, 0.980, and 0.974 in training, validation, and testing phases and low error values, especially in the case of NBMG, which

Table 3 The result of developed models for NB

Model	Phase	Index values				
		RMSE	R^2	MSE	MDAPE	NSE
NB	Train	1.742	0.968	3.035	6.137	0.966
	Validation	2.051	0.958	4.205	7.328	0.958
	Test	2.147	0.953	4.610	7.482	0.949
	All	1.856	0.963	3.446	6.442	0.962
NBMG	Train	1.129	0.986	1.275	2.914	0.986
	Validation	1.523	0.980	2.319	4.024	0.977
	Test	1.619	0.974	2.620	5.055	0.971
	All	1.278	0.983	1.633	3.237	0.982
NBHH	Train	1.428	0.978	2.039	2.555	0.977
	Validation	1.960	0.965	3.842	3.650	0.962
	Test	1.680	0.970	2.821	4.188	0.969
	All	1.557	0.975	2.426	3.012	0.973

are almost twice lower than NB single model indicate superior optimization performance of MGO in enhancing CL prediction capability of NB.

- *NBHH (NB + HHO)*: This model with marginal lower R^2 (lower than 1%) and higher error values (on average 20%) has weaker performance than NBMG. However, the MGO algorithm has notably enhanced the NB's prediction accuracy.

Figure 2 visually illustrates the trends in error values (RMSE, MSE) and R^2 for the three models developed in this study. The comparative analysis reveals a consistent decrease in R^2 values from training to testing across all models, indicating a weakness in the training ability of the models. Notably, all data columns for R^2 values of NBMG are higher than those of NB but show similar heights to NBHH. In terms of RMSE and MSE error values, the NBMG model, particularly during the training phase, demonstrated significantly lower error values compared to the other models. As detailed in Table 3 and depicted in Fig. 2, the NBMG model showcased the best performance in predicting CL values, boasting an impressive R^2 of 0.986, RMSE of 1.129 KW, and MSE of 1.275 KW.

Figure 3 provides a comprehensive visual representation through a scatter plot, elucidating the relationship between predicted and measured samples for the CL. The scrutiny of these samples unfolds across three distinct phases, each phase offering valuable insights into the model's performance. The allocation of sample points in the plot is guided by two main metrics: RMSE, which characterizes the dispersion within the figure, and R^2, a measure that assesses the degree of collinearity among the sample points. In this visual exploration, the coincidence of a high R^2 value with a low RMSE value signifies an optimal state where the predicted values closely align with the measured values, approximating the center ($X = Y$). To facilitate interpretation, two dashed lines are introduced onto the plot, delineating 15% overestimation and underestimation. Significantly, upon closer examination, the NBMG and NBHH hybrid models emerge as standout performers. These models, marked by their lowest RMSE values and highest R^2 values, showcase a level of performance that surpasses the NB single model. It is worth highlighting that while the NBHH model exhibits

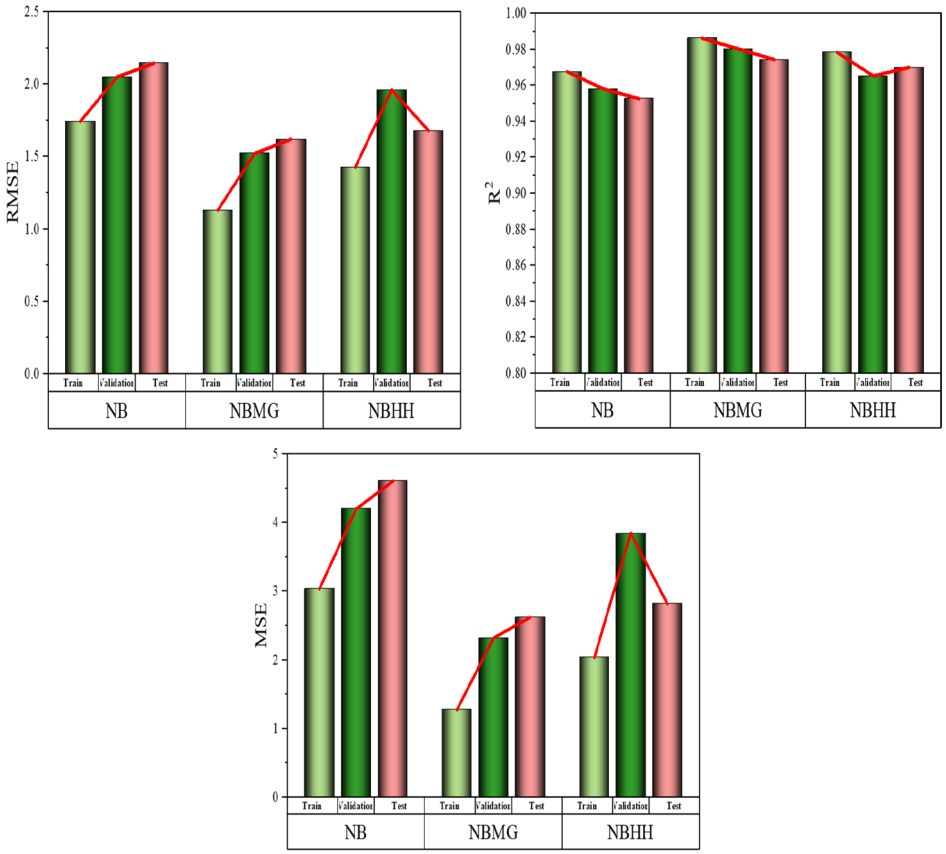

Fig. 2 The comparison of parameters

some comparative weakness against the NBMG model, it does present certain data points with overestimation exceeding 15%. This nuanced observation adds depth to the understanding of the models' performance dynamics across various scenarios and contributes to a more comprehensive evaluation of their predictive capabilities.

Figure 4 employs a line plot in this investigation to comprehensively compare the variation in error values across three developed models. The range of errors for NBMG is approximately half that of NBHH, underscoring the advantageous capability of the MGO algorithm. Furthermore, in the case of NBMG, the error rate during the training phase is only half that observed in the other two phases, suggesting that MGO exhibits superior prediction performance during the training phase compared to the other models. This observation is corroborated by Fig. 5, which illustrates the normal distribution of errors for MGO, displaying a narrow bell-shaped curve indicative of a high concentration of errors near 0%.

Figure 6 presents Taylor diagrams that vividly depict the performance of the employed predictive models, namely NB, NBMG, and NBHH. These diagrams serve as statistical syntheses, integrating both observed and predicted CL and incorporating essential metrics such as RMSE, correlation coefficients (CC), and normalized standard deviations. The visual representation within the figure provides a comprehensive overview of the model performances. Notably, the NBMG model, an amalgamation of the NB model, and the MGO optimizer emerge as the optimal predictive model. The outcomes of this

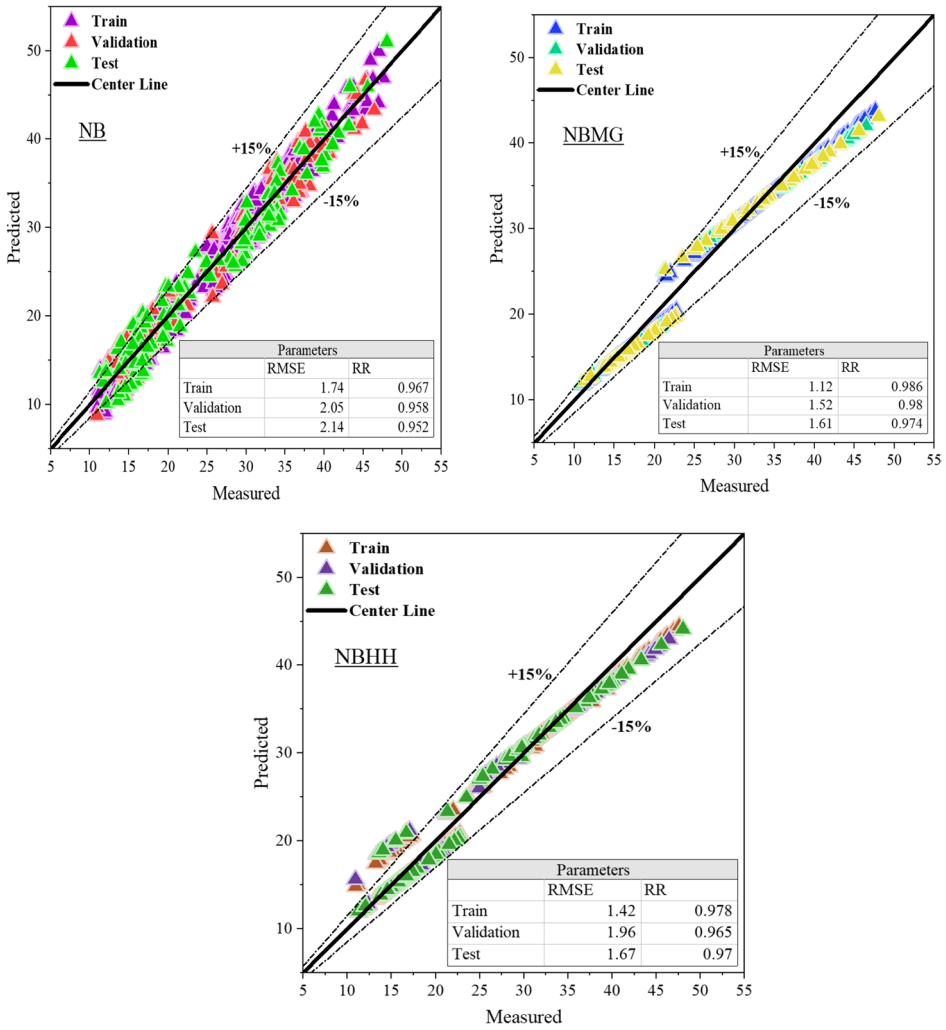

Fig. 3 The scatter plot for developed hybrid models

model closely align with the ideal benchmark observed in the experimental data. This alignment signifies the effectiveness of the NBMG model in capturing the intricate patterns of the cooling load, emphasizing its superior predictive capabilities compared to the other models under consideration.

Examining the kernel smooth distribution of errors during the prediction of CL values across the training, validation, and testing phases, Fig. 7 provides a graphical insight into the performance of three distinct models (NB, NBMG, and NBHH). Notably, the NB model displayed the highest errors during the testing phase, whereas the NBMG model showcased the lowest errors. Consistent favorability toward the NBMG hybrid model emerged across all stages of analysis. In the testing phase of the NB model, errors ranged widely from − 25 to 30. Conversely, the NBMG model, exhibiting superior performance during the training phase, featured errors predominantly concentrated within a narrower range of − 15 to 15. This emphasis on a refined error distribution underscores the

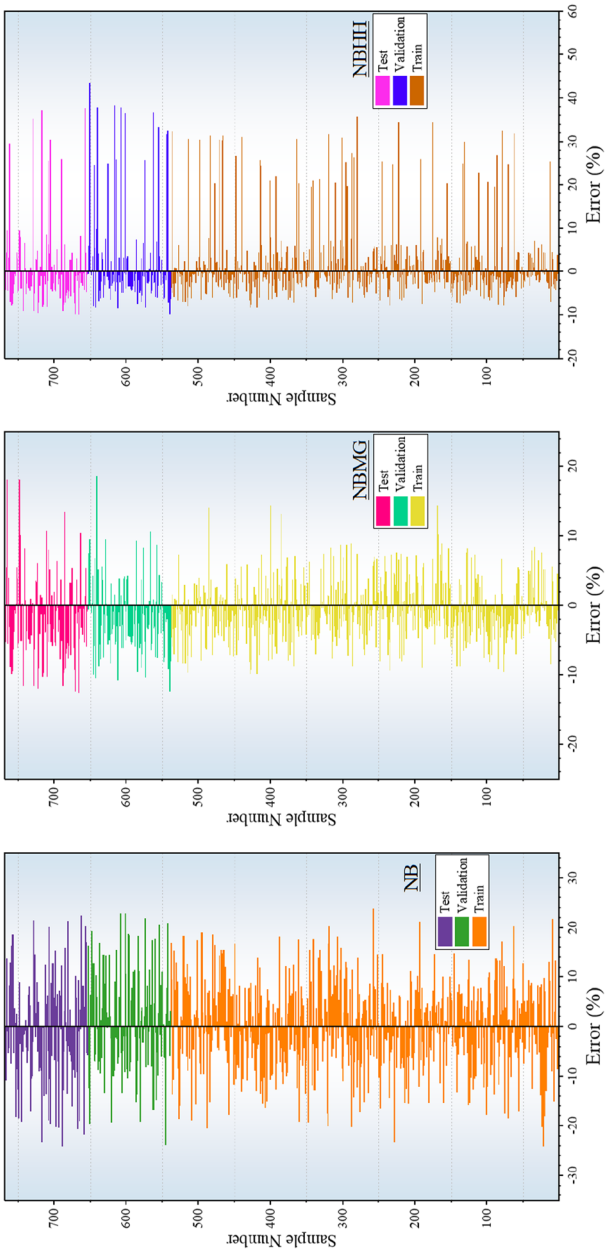

Fig. 4 The error rate percentage for the hybrid models is based on the line plot

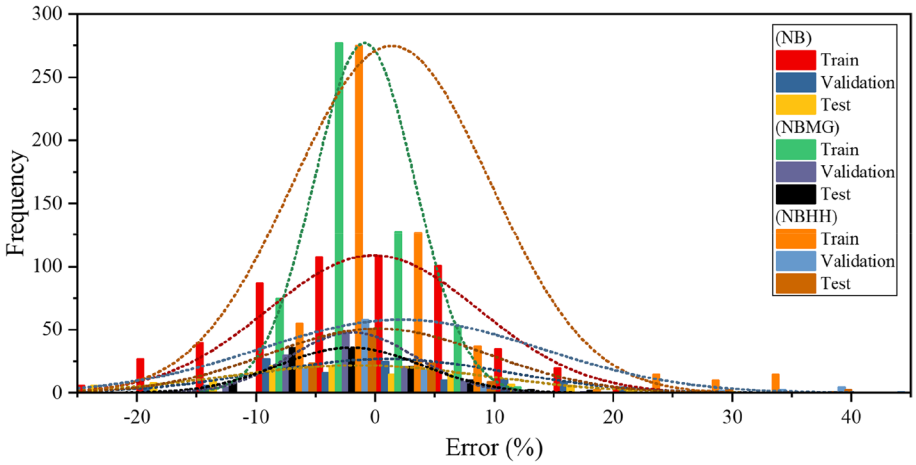

Fig. 5 The normal distribution plot of errors among the developed models

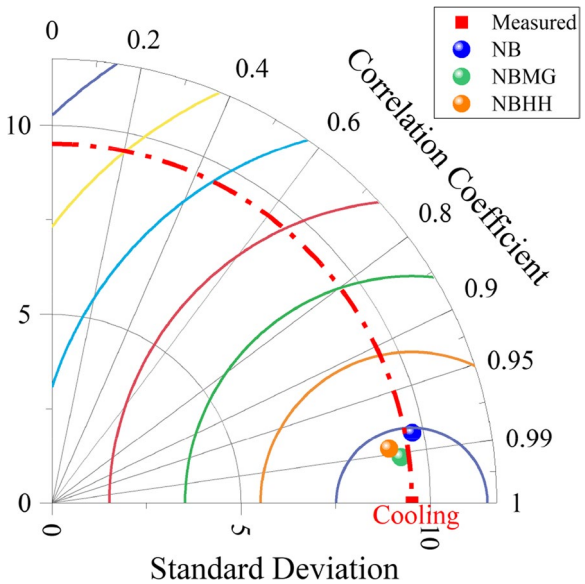

Fig. 6 The Taylor diagram for developed models

heightened predictive accuracy of the NBMG model, especially when compared to the broader range observed in the NB model's testing phase.

Conclusions

Accurate building cooling load forecasting is vital for optimizing HVAC systems, reducing costs, and enhancing energy efficiency. However, it remains challenging due to the complex interplay of building characteristics and meteorological data. Prior studies emphasize the effectiveness of machine learning in building energy forecasting, favoring nonlinear approaches. Naive Bayes, a foundational machine learning algorithm, was

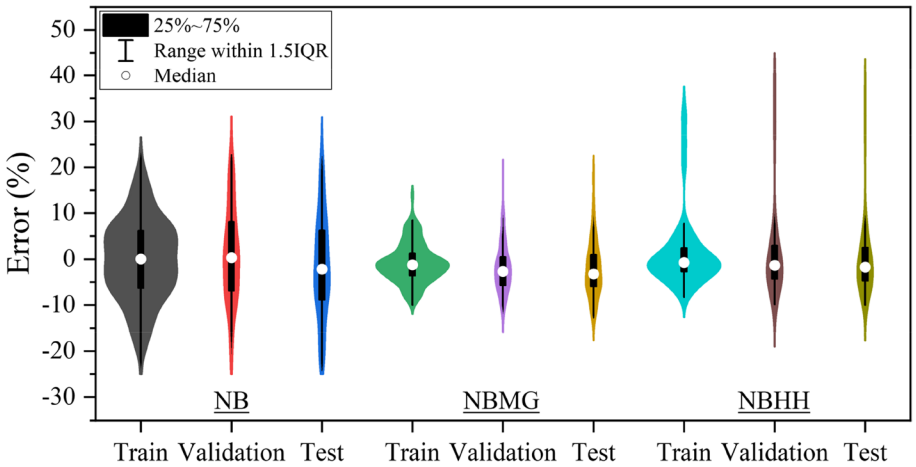

Fig. 7 The kernel smooth plot of errors among the developed models

unexplored in this context. Naive Bayes-based models encompassed a single model, one optimized with the Mountain Gazelle Optimizer (MGO) and another optimized with the horse herd optimization (HHO) algorithm. The research findings underscore the exceptional performance of the NBMG model, consistently outperforming its counterparts by reducing prediction errors by an average of 20% and achieving a maximum R^2 value of 0.982 for cooling load prediction. This highlights the substantial potential of machine learning, as NBMG exemplifies, to significantly enhance the precision of energy consumption forecasts. Consequently, it empowers decision-makers in energy conservation and retrofit strategies, contributing to the overarching goals of sustainable building operations and reduced environmental impact.

Acknowledgements
I would like to take this opportunity to acknowledge that there are no individuals or organizations that require acknowledgment for their contributions to this work.

Authors' contributions
The author contributed to the study's conception and design. Data collection, simulation, and analysis were performed by "YX."

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of data and materials
Data can be shared upon request.

Declarations

Competing interests
The author declares no competing interests.

Received: 3 December 2023 Accepted: 21 February 2024
Published online: 20 March 2024

References

1. Leitao J, Gil P, Ribeiro B, Cardoso A (2020) A survey on home energy management. IEEE Access 8:5699–5722
2. Gong H, Rallabandi V, McIntyre ML, Hossain E, Ionel DM (2021) Peak reduction and long term load forecasting for large residential communities including smart homes with energy storage. IEEE Access 9:19345–19355

3. Hannan MA, Faisal M, Ker PJ, Mun LH, Parvin K, Mahlia TMI, Blaabjerg F (2018) A review of Internet of energy based building energy management systems: issues and recommendations, Ieee. Access 6:38997–39014
4. Sadeghian O, Moradzadeh A, Mohammadi-Ivatloo B, Abapour M, Anvari-Moghaddam A, Lim JS, Marquez FPG (2021) A comprehensive review on energy saving options and saving potential in low voltage electricity distribution networks: building and public lighting. Sustain Cities Soc 72:103064
5. Sadeghian O, Moradzadeh A, Mohammadi-Ivatloo B, Abapour M, Garcia Marquez FP (2020) Generation units maintenance in combined heat and power integrated systems using the mixed integer quadratic programming approach. Energies (Basel). 13:2840
6. Nami H, Anvari-Moghaddam A, Arabkoohsar A (2020) Application of CCHPs in a centralized domestic heating, cooling and power network—thermodynamic and economic implications. Sustain Cities Soc 60:102151
7. Chen Q, Xia M, Lu T, Jiang X, Liu W, Sun Q (2019) Short-term load forecasting based on deep learning for end-user transformer subject to volatile electric heating loads. IEEE Access 7:162697–162707
8. Yao Y, Lian Z, Liu S, Hou Z (2004) Hourly cooling load prediction by a combined forecasting model based on analytic hierarchy process. Int J Therm Sci 43:1107–1118
9. Probst O (2004) Cooling load of buildings and code compliance. Appl Energy 77:171–186
10. Bojić M, Yik F (2005) Cooling energy evaluation for high-rise residential buildings in Hong Kong. Energy Build 37:345–351
11. Ansari FA, Mokhtar AS, Abbas KA, Adam NM (2005) A simple approach for building cooling load estimation. Am J Environ Sci 1:209–212
12. Shahzad MW, Burhan M, Ybyraiymkul D, Oh SJ, Ng KC (2019) An improved indirect evaporative cooler experimental investigation. Appl Energy. 256:113934. https://doi.org/10.1016/j.apenergy.2019.113934
13. Moradzadeh A, Moayyed H, Zakeri S, Mohammadi-Ivatloo B, Aguiar AP (2021) Deep learning-assisted short-term load forecasting for sustainable management of energy in microgrid. Inventions 6:15
14. Chen S, Zhang X, Wei S, Yang T, Guan J, Yang W, Qu L, Xu Y (2019) An energy planning oriented method for analyzing spatial-temporal characteristics of electric loads for heating/cooling in district buildings with a case study of one university campus. Sustain Cities Soc 51:101629
15. Tsanas A, Goulermas JY, Vartela V, Tsiapras D, Theodorakis G, Fisher AC, Sfirakis P (2009) The Windkessel model revisited: a qualitative analysis of the circulatory system. Med Eng Phys 31:581–588
16. Chou J-S, Bui D-K (2014) Modeling heating and cooling loads by artificial intelligence for energy-efficient building design. Energy Build 82:437–446
17. Bojic M, Yik F, Wan K, Burnett J (2000) Investigations of cooling loads in high-rise residential buildings in Hong Kong, in: Thermal Sciences 2000. Proceedings of the International Thermal Science Seminar. Volume 1, Begel House Inc
18. Chou SK, Chang WL (1997) Large building cooling load and energy use estimation. Int J Energy Res 21:169–183
19. Sodha MS, Kaur B, Kumar A, Bansal NK (1986) Comparison of the admittance and Fourier methods for predicting heating/cooling loads. Sol Energy (United Kingdom) 36
20. Mui KW, Wong LT (2007) Cooling load calculations in subtropical climate. Build Environ 42:2498–2504
21. Shin M, Do SL (2016) Prediction of cooling energy use in buildings using an enthalpy-based cooling degree days method in a hot and humid climate. Energy Build 110:57–70
22. Yun K, Luck R, Mago PJ, Cho H (2012) Building hourly thermal load prediction using an indexed ARX model. Energy Build 54:225–233
23. Korolija I, Zhang Y, Marjanovic-Halburd L, Hanby VI (2013) Regression models for predicting UK office building energy consumption from heating and cooling demands. Energy Build 59:214–227
24. Deb C, Eang LS, Yang J, Santamouris M (2016) Forecasting diurnal cooling energy load for institutional buildings using artificial neural networks. Energy Build 121:284–297
25. Gunay B, Shen W, Newsham G (2017) Inverse blackbox modeling of the heating and cooling load in office buildings. Energy Build 142:200–210
26. Kavaklioglu K (2011) Modeling and prediction of Turkey's electricity consumption using support vector regression. Appl Energy 88:368–375
27. Li Q, Meng Q, Cai J, Yoshino H, Mochida A (2009) Applying support vector machine to predict hourly cooling load in the building. Appl Energy 86:2249–2256
28. Moradzadeh A, Sadeghian O, Pourhossein K, Mohammadi-Ivatloo B, Anvari-Moghaddam A (2020) Improving residential load disaggregation for sustainable development of energy via principal component analysis. Sustainability 12:3158
29. Zhao J, Liu X (2018) A hybrid method of dynamic cooling and heating load forecasting for office buildings based on artificial intelligence and regression analysis. Energy Build 174:293–308
30. Roy SS, Roy R, Balas VE (2018) Estimating heating load in buildings using multivariate adaptive regression splines, extreme learning machine, a hybrid model of MARS and ELM. Renew Sustain Energy Rev 82:4256–4268
31. Moradzadeh A, Zeinal-Kheiri S, Mohammadi-Ivatloo B, Abapour M, Anvari-Moghaddam A (2020) Support vector machine-assisted improvement residential load disaggregation, in: 2020 28th Iranian Conference on Electrical Engineering (ICEE). IEEE 1–6
32. Luo XJ, Oyedele LO, Ajayi AO, Akinade OO (2020) Comparative study of machine learning-based multi-objective prediction framework for multiple building energy loads. Sustain Cities Soc 61:102283
33. Moradzadeh A, Zakeri S, Shoaran M, Mohammadi-Ivatloo B, Mohammadi F (2020) Short-term load forecasting of microgrid via hybrid support vector regression and long short-term memory algorithms. Sustainability 12:7076
34. Ding Y, Su H, Kong X, Zhang Z (2020) Ultra-short-term building cooling load prediction model based on feature set construction and ensemble machine learning. IEEE Access 8:178733–178745
35. Wang Z, Hong T, Piette MA (2019) Data fusion in predicting internal heat gains for office buildings through a deep learning approach. Appl Energy 240:386–398
36. Roy SS, Samui P, Nagtode I, Jain H, Shivaramakrishnan V, Mohammadi-Ivatloo B (2020) Forecasting heating and cooling loads of buildings: a comparative performance analysis. J Ambient Intell Humaniz Comput 11:1253–1264

37. Song J, Xue G, Pan X, Ma Y, Li H (2020) Hourly heat load prediction model based on temporal convolutional neural network. IEEE Access 8:16726–16741
38. Yu Z, Haghighat F, Fung BCM, Yoshino H (2010) A decision tree method for building energy demand modeling. Energy Build 42:1637–1646
39. Ahmad T, Chen H (2018) Short and medium-term forecasting of cooling and heating load demand in building environment with data-mining based approaches. Energy Build 166:460–476
40. Moradzadeh A, Mansour-Saatloo A, Mohammadi-Ivatloo B, Anvari-Moghaddam A (2020) Performance evaluation of two machine learning techniques in heating and cooling loads forecasting of residential buildings. Appl Sci 10:3829
41. Geysen D, De Somer O, Johansson C, Brage J, Vanhoudt D (2018) Operational thermal load forecasting in district heating networks using machine learning and expert advice. Energy Build 162:144–153
42. Cui B, Fan C, Munk J, Mao N, Xiao F, Dong J, Kuruganti T (2019) A hybrid building thermal modeling approach for predicting temperatures in typical, detached, two-story houses. Appl Energy 236:101–116
43. Wang R, Lu S, Feng W (2020) A novel improved model for building energy consumption prediction based on model integration. Appl Energy 262:114561
44. Chen Q, M Kum Ja, Burhan M, Akhtar FH, Shahzad MW, Ybyraiymkul D, Ng KC (2021) A hybrid indirect evaporative cooling-mechanical vapor compression process for energy-efficient air conditioning. Energy Convers Manag. 248:114798. https://doi.org/10.1016/j.enconman.2021.114798
45. Wong SL, Wan KKW, Lam TNT (2010) Artificial neural networks for energy analysis of office buildings with daylighting. Appl Energy 87:551–557
46. Paudel S, Elmtiri M, Kling WL, Le Corre O, Lacarrière B (2014) Pseudo dynamic transitional modeling of building heating energy demand using artificial neural network. Energy Build 70:81–93
47. Schiavon S, Lee KH, Bauman F, Webster T (2010) Influence of raised floor on zone design cooling load in commercial buildings. Energy Build 42:1182–1191
48. Fan C, Wang J, Gang W, Li S (2019) Assessment of deep recurrent neural network-based strategies for short-term building energy predictions. Appl Energy 236:700–710
49. Zhong H, Wang J, Jia H, Mu Y, Lv S (2019) Vector field-based support vector regression for building energy consumption prediction. Appl Energy 242:403–414
50. B.S.A.J. khiavi; B.N.E.K.A.R.T.K. hadi Sadaghat (2023) The utilization of a Naïve Bayes model for predicting the energy consumption of buildings. J Art Intel Syst Modelling 01. https://doi.org/10.22034/JAISM.2023.422292.1003
51. Zhou G, Moayedi H, Bahiraei M, Lyu Z (2020) Employing artificial bee colony and particle swarm techniques for optimizing a neural network in prediction of heating and cooling loads of residential buildings. J Clean Prod 254:120082
52. Pessenlehner W, Mahdavi A (2023) Building morphology, transparence, and energy performance, na
53. Hastie T, Tibshirani R, Friedman JH, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, Springer, New York City
54. Piryonesi SM, El-Diraby TE (2020) Role of data analytics in infrastructure asset management: overcoming data size and quality problems. J Transportation Eng Part B: Pavements 146:4020022
55. Abdollahzadeh B, Gharehchopogh FS, Khodadadi N, Mirjalili S (2022) Mountain gazelle optimizer: a new nature-inspired metaheuristic algorithm for global optimization problems. Adv Eng Softw 174:103282
56. MiarNaeimi F, Azizyan G, Rashki M (2021) Horse herd optimization algorithm: a nature-inspired algorithm for high-dimensional optimization problems. Knowl Based Syst 213:106711

Publisher's Note