Novel Data-Driven Machine Learning Models for Heating Load Prediction: Single and Optimized Naive Bayes

Fangyuan Li*
Zhejiang Business Technology Institute, Ningbo, Zhejiang, 315012, China

***Abstract*—Numerous approaches can be employed to create models for assessing the heat gains of a building arising from both external and internal sources. This modeling process evaluates effective operational strategies, conducts retrofit audits, and projects energy consumption. These techniques range from simple regression analyses to more intricate models grounded in physical principles. A prevalent assumption underlying all these modeling techniques is the requirement for input variables to be derived from authentic data, as the absence of realistic input data can lead to substantial underestimations or overestimations in energy consumption assessments. In this paper, eight input parameters, including relative compactness, orientation, wall area, roof area, glazing area, overall height, surface area, and glazing area distribution, are employed for training proposed Naive Bayes (NB)-based machine learning models. Utilizing a novel approach, this research explores the application of Beluga Whale Optimization and the Coot Optimization algorithm for optimizing the Naive Bayes model in heating load prediction. By harnessing the collective intelligence of Beluga Whales and drawing from the cooperative behavior of coots, the research aims to improve the model's predictive capabilities, which is of paramount importance in building energy management. Based on the comparative analysis between developed models (NB, NBCO, and NBBW), it is attainable that NBCO and NBBW, as two optimized models, have 2.4% and 1.3% higher R^2 values, respectively. Also, the RMSE of the NBCO was, on average, 19-33% lower than that of the two other models, confirming the high accuracy of NBCO. This innovative integration of bio-inspired optimization techniques with machine learning demonstrates a promising avenue for optimizing predictive models, offering potential energy efficiency and sustainability advancements in the built environment.**

Keywords*—Prediction models; heating load demand; building energy consumption; Naive Bayes; metaheuristic optimization algorithms

I. Introduction

In contemporary facility management, a critical challenge managers face revolves around assessing and predicting a building's energy requirements, particularly those equipped with air conditioning systems. This challenge stems from the notable variability exhibited in the energy feeding patterns generated by these systems. These fluctuations can be attributed to changes in external climate situations, the ebb and flow of occupants throughout the day, and internal loads incorporated within the building [1]. A holistic understanding of building performance is imperative to address this challenge and optimize building energy consumption. This begins with the initial identification of energy resources and the principal end-uses within the building. Energy resources typically natural gas, encompass electricity, and district heating supply, while the major end-uses comprise heating, ventilation, and air-conditioning ($HVAC$) systems, domestic hot water, lighting, plug-loads, elevators, kitchen equipment, ancillary appliances, and various equipment [2]. Such an integrated approach can enhance energy management and sustainability in the built environment.

Scholars have developed diverse assessment systems and modeling methodologies to propose an optimal predictive tool for estimating building energy consumption [3, 4]. Within this framework, two conventional devices for evaluating the Energy Performance of Buildings (EPB) through modeling and simulation have been highlighted in lectures [5]. Historically, the physical attributes of buildings, such as their geometry, were used as the basis for energy performance simulations. However, using such predictors has limitations, as it entails controlling factors that are challenging to manipulate in practical applications [6]. This can be considered a drawback of these traditional approaches. The primary challenge in advancing simulation and modeling techniques is the accurate estimation of EPB, a time-consuming process demanding meticulous attention due to including many influencing factors.

Furthermore, using various simulation programs may yield assessments with varying degrees of accuracy [7]. These methods can be relied upon to calculate the impact of individual factors on EPB when all other variables remain constant. Notably, there are established computer software tools like Designer's Simulation Toolkit (DeST) [8], Energy Plus [9], and DOE-2 [10] that facilitate these modeling and simulation endeavors.

Engineers have proposed using inverse (data-driven) modeling to remedy the limitations associated with simulation tools to explore EPB [11]. In this approach, a robust assessment of the impact of significant factors (e.g., roof area, relative compactness, and orientation) on EPB can be achieved by ensuring a sufficient quantity of data samples [12–15]. Various Machine Learning (ML) models, due to their ease of implementation and high-performance speed [16–18], were highly regarded by scholars.

For instance, Kalogirou and Bojic [19] employed a recurrent neural network to predict the energy feasting of a passive solar building. Pao [20] compared various models and concluded that ANN models are well-suited for forecasting building energy

consumption, effectively capturing complex non-linear relationships. Ben – Nakhi and Mahmoud [21] used ANN models to predict building cooling loads, achieving a strong fit to experimental data and optimizing thermal energy storage in public and office buildings.

In addition to Artificial Neural Networks ($ANNs$), various other artificial intelligence (AI) tools, including Support Vector Machine (SVM) [22] regression, neuro–fuzzy systems [23], and random forests [24], have been applied to address EPB challenges. For instance, Li et al. [25] conducted a qualified study on cooling load calculations, demonstrating the effectiveness of SVM and General Regression Neural Network (GRNN) compared to conventional ANNs. Moreover, researchers [26] have integrated SVM and wavelet transforms with Partial Least Squares Regression (PLS) to model office building heating and cooling loads, yielding precise insights. While AI has proven valuable in EPB, computational challenges have prompted the use of metaheuristic algorithms like Genetic Algorithm and Particle Swarm Optimization [27, 28, 38], which this study further explores, focusing on Beluga Whale Optimization (BWO) [29] and the Coot Optimization algorithm (COA) [30] for optimizing the Naive Bayes (NB) [31] model in heating load prediction.

The NB model is a widely used machine learning (ML) algorithm known for its simplicity and effectiveness in classification tasks in many applications similar to this study [32–34]. It is based on Bayes' theorem and chin independence assumption, making it particularly suited for applications where the independence assumption holds. It calculates the likelihood of a particular instance belonging to a specific class based on the probabilities of its features occurring in each class. This study embarks on developing NB-based models for predicting heating loads (HL) in buildings. Two distinct optimizers, as mentioned above (BWO and COA), were employed to optimize the training process. The predicted results of the $three$ models were subjected to comparison utilizing performance metrics, including R^2, RMSE, MSE, U95, and IOA. Afterward, the most optimal hybrid model for predicting HL in buildings was identified.

The choice of Naive Bayes (NB)--based machine learning models is particularly appropriate for addressing this type of problem due to several reasons. Firstly, NB models are known for their simplicity and computational efficiency, making them well-suited for handling large datasets and multiple input variables, such as those involved in predicting building heating loads. Secondly, NB models assume conditional independence between input features, which, despite being a simplification, often works well in practice, especially in complex systems where interactions between variables may not be easily discernible. This makes NB models robust and less prone to overfitting compared to more complex algorithms. Additionally, the probabilistic nature of NB models allows for clear interpretability of the results, providing insights into the contribution of each feature to the prediction, which is valuable in the situation of building energy management. Finally, the integration of bio-inspired optimization techniques like BWO and COA further enhances the model's ability to fine-tune its parameters, leading to improved accuracy and reliability in heating load predictions. This combination of simplicity, efficiency, and optimization makes NB-based models an effective choice for tackling the challenges of energy modeling in buildings. The paper is organized into five sections. The Abstract provides a concise summary of the study's objectives, methods, and key findings. The Introduction in Section I outlines the research background, related works, and significance. Materials and Methods in Section II details the dataset, machine learning models, and the hybrid optimization algorithms used, along with the evaluation metrics. The Results in Section III presents the outcomes of the modeling process. Discussion in Section IV offers a validation of present study, compares their performance, and addresses the study's limitations. Finally, the Conclusion in Section VI summarizes the findings, discusses implications for energy management, and suggests avenues for future research.

II. Materials and Methods

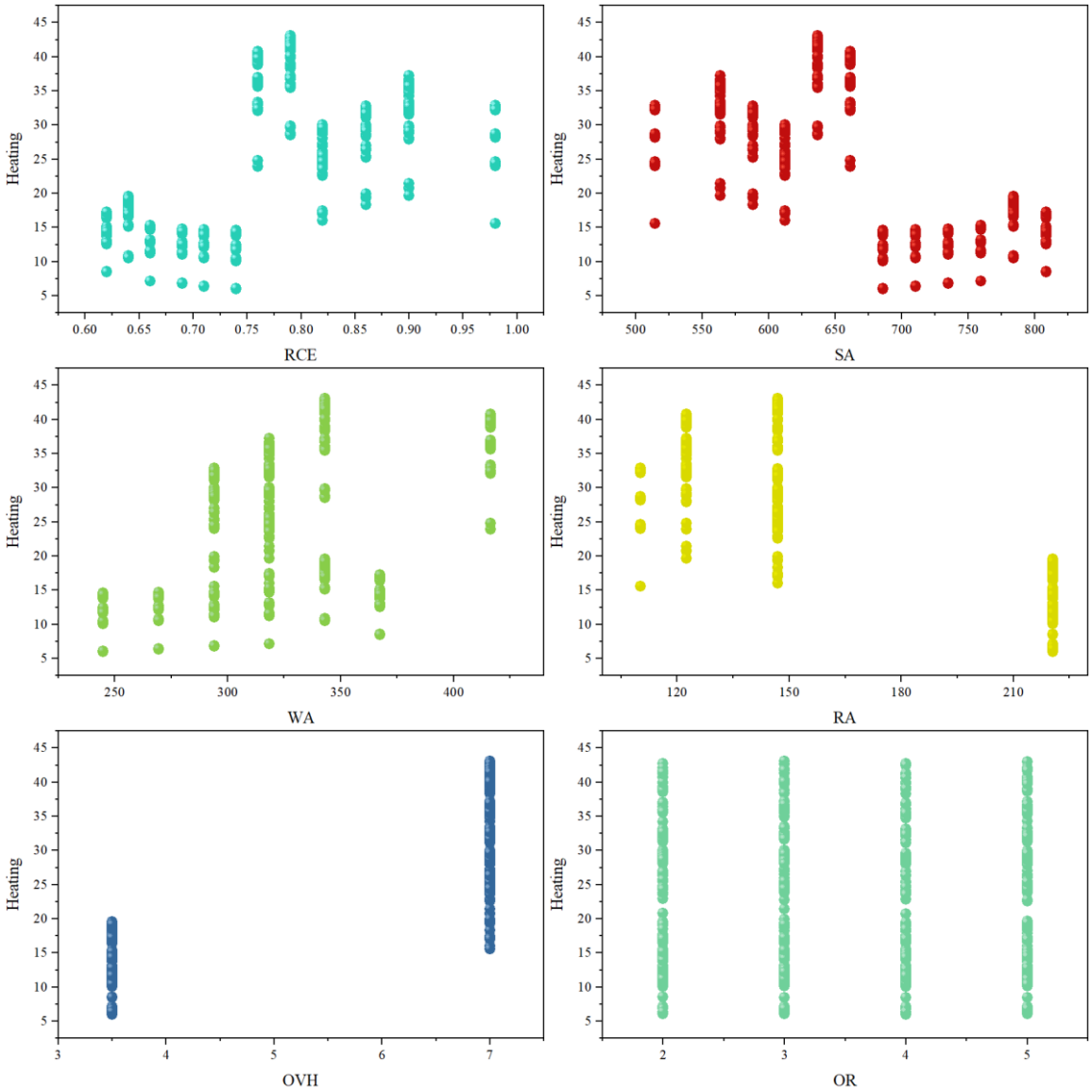
A. Dataset Description

The primary aim of this study is to predict HL in buildings by utilizing data that captures energy consumption patterns. A simulation approach involving the NB model is employed to accomplish this objective, with the training process incorporating two distinct optimizers designed to optimize NB hyper parameters. The inputs provided to the predictive model encompass various parameters, including relative compactness (RCE), $surface\ area\ \ (SA)$, $overall\ height\ \ (OVH)$, $roof\ area\ \ (RA)$, $glazing\ area\ \ (GA)$, $wall\ area\ \ (WA)$, $orientation\ (OR)$, and $glazing\ area\ distribution\ (GAD)$. The data relating to the input and output parameters, including minimum, maximum, average, standard deviation, Median, and Skewness, is reported in Table I. Minimum and maximum values identify the lowest and highest data points, establishing the data's range. The average, also known as the mean, provides a central measure to understand the typical value in the dataset. Standard deviation quantifies the dispersion of data points, indicating how closely they cluster around the mean. Conversely, the median represents the middle value when the data is ordered, making it robust to outliers. Skewness measures the asymmetry of the data distribution, indicating whether it is skewed to the left or right.

The scatter plot in Fig. 1 demonstrates the correlation among input and output parameters. The data distribution related to the RCE, SA, and WA input parameters is vertically highly asymmetric with the highest skewness values (RCE and WA skewed right of the average and SA skewed left of the average). Data points of OVH are located in two values (3.5 and 7) where $OVH = 3.5$ is related to lower heating values (below 20), and $OVH = 7$ corresponds to the heating values higher than 20. The OVH and OR data points' distribution is highly symmetric, with skewness values approximately equal to zero and their median and average values the same. GA and GAD are the only parameters with zero values, indicating that their effect is neglected in some samples.

TABLE I. THE STATISTICAL PROPERTIES OF THE INPUT ADJUSTABLE OF HEATING

Variables	*Category*	*Indicators*					
		Min	***Max***	***Median***	***Avg***	***Skew***	***St.Dev.***
RCE	*Input*	0.62	0.98	0.75	0.764	0.496	0.106
SA	*Input*	514.5	808.5	673.75	671.708	−0.125	88.09
WA	*Input*	245	416.5	318.5	318.5	0.534	43.63
RA	*Input*	110.25	220.5	176.604	176.604	−0.163	45.17
OVH	*Input*	3.5	7	5.25	5.25	$-2.9E-19$	1.751
OR	*Input*	2	5	3.5	3.5	$2.68E-18$	1.119
GA	*Input*	0	0.4	0.234	0.235	−0.060	0.133
GAD	*Input*	0	5	2.813	2.813	−0.089	1.551
Heating	*Output*	6.01	42.96	22.307	22.307	0.361	10.09



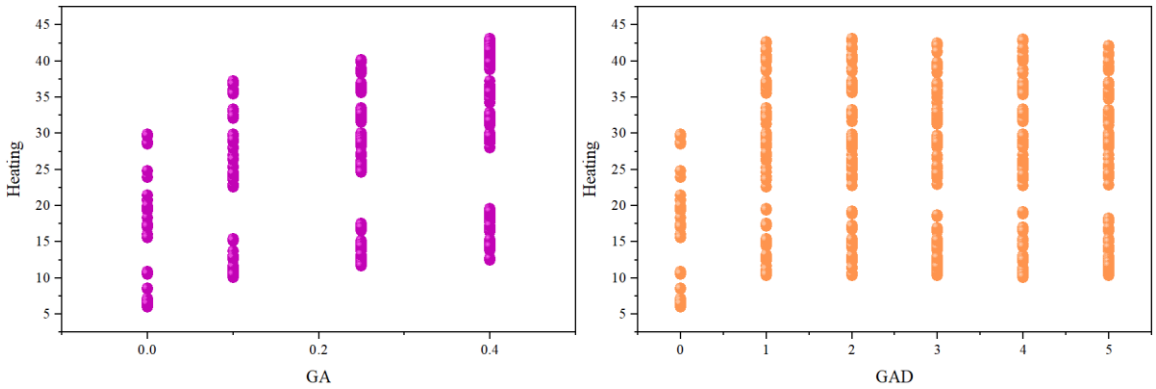


Fig. 1. Scatter plot amid input and output.

B. Machine Learning Model and Hybrid Optimization Algorithms

1) Naive bayes (NB): The NB is a robust probabilistic classification method grounded in Bayes' theorem, streamlining the modeling process by assuming input variable independence. When integrated with kernel density approximations, NB exhibits promise for significant enhancements in predictive accuracy, as indicated in previous studies [35, 36]. Notably, NB stands out due to its scalability, characterized by a need for only a few input parameters that increase linearly with the number of predictors. This differentiates it from computationally demanding classifiers. The closed-form training methodology of NB is remarkably efficient, ensuring swifter performance compared to more intricate computational techniques.

The NB classifier represents an advanced system seamlessly incorporating the NB probability model into its decision-making framework. Its foundation lies in applying the max a posteriori (MAP) choice rule, a proven approach for selecting the most likely supposition from a set of possible choices. Furthermore, it is worth noting the existence of a closely linked classifier known as the Bayes classifier. This formidable algorithm plays a pivotal role in assigning class labels $y = C_k$, with k ranging from 1 to K, a process involving an intricate assessment of multiple factors and variables to classify data points into predetermined categories.

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

2) Beluga whale optimization (BWO): The BWO method simulates beluga whale (BW) behaviors for optimization, with two phases: exploration and refinement, using beluga whales as search agents updating candidate solutions within a specified area. The matrix maps the positions of these search agents (Zhong, Li, und Meng 2022):

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,d} \\ x_{2,1} & x_{2,2} & x_{2,d} \\ x_{n,1} & x_{n,2} & x_{n,d} \end{bmatrix} \tag{2}$$

Within this framework, 'n' signifies the people size of BW, and $'d'$ denotes the dimensionality of design variables, with the fitness values of each individual within this population being meticulously recorded as follows:

$$F_X = \begin{bmatrix} f(x_{1,1}, x_{1,2}, \dots, x_{1,n}) \\ f(x_{2,1}, x_{2,2}, \dots, x_{2,d}) \\ f(x_{n,1}, x_{n,2}, \dots, x_{n,d} \end{bmatrix} \tag{3}$$

The changeover from examination to exploitation in the BWO algorithm is determined by the mathematical representation of the balance factor B_f.

$$B_f = B_0(1 - T/2T_{max}) \tag{4}$$

Throughout each iteration, random fluctuations within the range of $(0, 1)$ are experienced by the value of B_0, with the current iteration being denoted by T, and the max allowable number of iterations being represented by T_{max}. The exploration stage is initiated when the balance factor (B_f) surpasses the threshold of 0.5, while the exploitation stage is engaged when Bf is either less than or equal to 0.5. As the number of iterations (T) escalates, the variability in Bf is observed to diminish from the initial span of $(0, 1)$ to the narrower interval of $(0, 0.5)$. This transformation underscores a conspicuous alteration in the probability of transitioning between the exploitation and exploration stages, with the likelihood of entering the exploitation phase being augmented as the iteration count progressively increases.

- Exploration phase

The exploration phase in BWO is inspired by observed synchronized swimming behaviors of captive beluga whales, influencing the search agents' coordinates and subsequent position modifications.

$$\begin{cases} X_{i,j}^{T+1} = X_{i,pj}^{T} + (X_{r,p1}^{T} - X_{i,pj}^{T})(1 + r_1)sin(2\pi r_2), & j = even \\ X_{i,j}^{T+1} = X_{i,pj}^{T} + (X_{r,p1}^{T} - X_{r,pj}^{T})(1 + r_1)\cos(2\pi r_2), & j = \text{odd} \end{cases} \tag{5}$$

Within this equation, the current iteration count, denoted as T, establishes the framework. The expression $X_{i,j}^{T+1}$ represents the newly adjusted location for the $i-th$ beluga whale along the $j-th$ dimension. Concomitantly, pj (with j spanning from 1 to d) is symbolic of a value randomly selected from the $d-$dimensional space. Moreover, $X_{i,pj}^{T}$ designates the $i-th$ BW position along the pj dimension at iteration T Furthermore, both $X_{i,pj}^{T}$ and $X_{r,pj}^{T}$ serve to depict the prevailing positions of the $i-th$ BW and a stochastically chosen $r-th$ beluga whale, where r is selected randomly. Additionally, $r1$

and $r2$ are arbitrary values within the range of $(0, 1)$. It is of significance to note that the sine (sin) and cosine (cos) functions, applied to $(2\pi r_2)$, delineate the alignment of the mirrored BW fins toward the water's external. The selection of dimensions using odd or even numbers determines the reflection of synchronized or mirrored behaviors exhibited by BW during swimming or diving in the updated position. To enhance the stochastic components within the exploration stage, two random values identified as $r_1$and r_2, are utilized.

- Exploitation Phase

The exploitation stage in BWO is inspired by BW cooperative foraging and adaptive movement patterns, involving sharing positional information and coordination, utilizing the Levy flight strategy for convergence enhancement (Mantegna 1994), which has been integrated into the exploitation phase of BWO. It is postulated that these whales employ the Levy flight strategy for capturing prey, and this strategy is expressed mathematically as follows:

$$X_i^{T+1} = r_3 X_{best}^T - r_4 X_i^T + C_1 . L_F . (X_r^T - X_i^T) \quad (6)$$

In the context of the current iteration designated as "T," the following elements are encompassed: X_i, which serves as a representation of the current position of the $i-th$ beluga whale and "X_r," which represents the current position of a beluga whale that has been randomly selected. Furthermore, "X_i^{T+1}" denotes the updated position of the $i-th$ beluga whale, and "X_{best}" designates the optimal position among all the beluga whales. Additionally, "r3" and "r4" signify randomly generated numbers that fall from 0 to 1. Lastly, "C_1" is ascertained utilizing a calculation involving "r_4", specifically it determines the value of r_4 multiplied by the expression " $C_1 = 2r_4(1 - T/T_{max})$ thereby representing the random jump strength that quantifies the magnitude of a Levy flight [37].

The Levy flight function, denoted as LF, is computed according to the following procedure.

$$L_F = 00.5 \times \frac{u \times \sigma}{|v|^{1/\beta}} \quad (7)$$

$$\sigma = (\frac{\Gamma(1+\beta) \times \sin(\pi\beta/2)}{\Gamma((1+\beta)/2) \times \beta \times 2^{(\beta-1)/2}} \quad (8)$$

In this context, β, the default constant set to 1.5, is accompanied by normally distributed random numbers u and v.

- Whale fall

In BWO iterations, whale falls are simulated to mimic the beluga whale population changes. Assuming that some whales relocate or descend to the ocean floor, positions and step magnitudes are adjusted to maintain population size, resembling the natural process of whale fall decomposition.

$$X_i^{T+1} = r_5 X_i^T - r_6 X_r^T + r_7 X_{step} \quad (9)$$

"X_{step} is the step size of whale fall, which is determined as follows: where r_5, r_6, and r_7 are random numbers within the range of $(0, 1)$."

$$X_{step} = (u_b - l_b)\exp(-C_2 T/T_{max}) \quad (10)$$

In this context, the parameterC_2 is possessed, functioning as the step factor and being linked to the probability of a whale fall event, along with the population size $(C_2 = 2W_f \times n)$. Furthermore, the variables u_b and l_b are present, signifying the $upper$ and $lower$ boundaries of variables, respectively. It is observable that the extent of the step size is influenced by a range of factors, encompassing the constraints established by the design variables, the ongoing iteration, and the max permissible number of iterations.

This model calculates the probability of a whale falling (W_f) as a linear function:

$$W_f = 0.1 - 0.05T/T_{max} \quad (11)$$

The decrease in the probability of a whale falling from 0.1 in the initial iteration to 0.05 in the final iteration indicates a trend in which, as the food source is approached more closely by beluga whales during the optimization process, the risk to beluga whales is mitigated.

3) Coot optimization algorithm (COA): The COA is influenced by the group behaviors of Coots, a water bird species, and utilizes a metaheuristic optimization strategy. Coots exhibit various movements on water as they seek food sources or specific destinations, including chain, random, leader – driven, and leader – adjusted motions. The COOT algorithm integrates these behaviors into its structure. In its application, the algorithm commences by randomly establishing a population, following the guidelines of Eq. (12) as specified in (Naruei und Keynia 2021):

$$CootPos(i) = rand\ (1, N) \times (UB - LB) + LB \quad (12)$$

$CootPos(i)$ signifies the geographical coordinates of an individual Coot, where N matches the dimensionality of issues or the count of involved variables. UB and LB, on the other hand, represent the $upper$ and $lower$ confines of the search space in which the pursuit is performed.

$$UB = [UB_1, UB_2, \dots, UB_N], LB = [LB_1, LB_2, \dots, LB_N] \quad (13)$$

After the initial population setup, four different crusade designs are used to adjust the coots' situations.

- Random Movement

Following the equation described in Eq. (14) below, position Q is first randomized for this particular movement:

$$Q = rand(1, N) \times (UB - LB) + LB \quad (14)$$

To avoid becoming trapped in local optima, the position is updated in line with the equation presented in Eq. (15):

$$CootPos(i) = CootPos(i) + A \times R_2 \times (Q - CootPos(i)) \quad (15)$$

To determine A, the R_2 is a random number that exists within the range $[0, 1]$, and its value is calculated by an equation given in Eq. (16):

$$A = 1 - L \times (\frac{1}{Iter}) \quad (16)$$

In this case, $Iter$ is the highest achievable number of iterations, and L is a reference to currently recorded numbers.

- Chain Movement

The regular location of two chick birds may be calculated by applying the formula in Eq. (17) to execute chain movements.

$$CootPos(i) = \frac{CootPos(i-1)+CootPos(i)}{2} \quad (17)$$

In this case, $CootPos(i-1)$ indicates the placement of another coot in the arrangement.

- Adjusting Location Giving to the Leader

A coot bird's place in each group is adjusted according to the leader's location, which causes the follower to move closer to the leader. The method given in Eq. (18) is used to estimate the leader's designation:

$$K = 1 + (i \ MOD \ NL) \quad (18)$$

K represents the index of leaders, i " refers to the supporter coot bird's sequence, and NL indicates the total total of leaders in the group.

Following the formula in Eq. (19), a coot's position will be modified during this specific motion.

$$CootPos(i) = LeaderPos(K) + 2 \times R_1 \times Cos(2R\pi) \times (LeaderPos(K) - CootPos(i)) \quad (19)$$

$CootPos(i)$ refers to the present location of the coot bird, and $LeaderPos(K)$ stands for the chosen leader's position. R_1 is a randomly generated number within the range of $[0, 1]$, and R is another random number within the range of $[-1, 1]$. These parameters are utilized in the location update calculation outlined in Eq. (19).

- Leader Movement

The leadership roles experience modifications as outlined in Eq. (20), aiming to shift from locally optimal positions to globally optimal ones.

$$LeaderPos(i) = \begin{cases} B \times B_3 \times Cos(2\pi R) \times \big(gBest - LeaderPos(i)\big) + gBest & B_4 < 0.5 \\ B \times B_3 \times Cos(2\pi R) \times \big(gBest - LeaderPos(i)\big) - gBest & B_4 \geq 0.5 \end{cases} \quad (20)$$

In this specific scenario, $gBest$ characterizes the optimal possible site, and B_3 and B_4 are haphazardly generated figures within the break $[0, 1]$. The value B is calculated using the equation provided in Eq. (21):

$$B = 2 - L \times (\frac{1}{Iter}) \quad (21)$$

a) Performance evaluation metrics: In the evaluation of the performance of a regression model, it is customary to employ the following metrics:

- Coefficient of Determination (R^2): Commonly represented as R^2 , measures the percentage of inconsistency in the reliant on variable that can be attributed to the sovereign variables within a statistical model. The following formula demonstrates it:

$$R^2 = \left(\frac{\sum_{i=1}^{n}(t_i-\bar{w})(v_i-\bar{v})}{\sqrt{[\sum_{i=1}^{n}(v_i-w)^2][\sum_{i=1}^{n}(v_i-\bar{v})^2]}} \right)^2 \quad (22)$$

- Error evaluation metrics (RMSE, MSE): $RMSE$ (Root Mean Square Error) and MSE (Mean Square Error) are statistical metrics that quantify the average magnitude and accuracy of errors among predicted and observed values in a model, with RMSE emphasizing the root of the squared differences. These metrics are mathematically represented in Eq. (23) and (24) as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_i - w_i)^2} \quad (23)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(v_i - w_i)^2 \quad (24)$$

- Uncertainty 95% (U95): This metric illustrates the range in which 95% of the predicted values correspond to the actual observed values. It provides valuable insights into the correctness and dependability of a model's predictions, particularly when evaluating its variation and level of uncertainty. The mathematical expression of this metric can be found in Eq. (25).

$$U95 = \sqrt{\sum_{i=1}^{n}(v_i - \bar{v})^2/(n*(n-1))} \quad (25)$$

- Index of Agreement (IOA): IOA is a metric used to evaluate the agreement or accuracy of model predictions compared to observed data, typically expressed as a value between 0 and 1. The Eq. (26) represents it below:

$$IOA = 1 - \frac{\sum_{i=1}^{n}|w_i-v_i|}{\sum_{i=1}^{n}(|w_i-\bar{w}|+|v_i-\bar{w}|)} \quad (26)$$

In all equations:

n: quantity of samples,

v_i: denotes the individual predicted cost,

$\bar{v}$: indicates the mean of the predicted morals,

w_i: stands for the experimentally measured cost,

$\bar{w}$: represents the average of the experimentally measured values.

III. Results

In this research paper, the assessment of heating energy consumption relies on utilizing a Naive Bayes (NB) model. Two optimization algorithms, COA and BWO, have been employed to assess the model's performance and training procedure. To create the requisite datasets for training, validation, and testing, a partitioning scheme of 70% for training, 15% for validation, and 15% for testing has been implemented.

In Table II, it is evident that the R^2 values exhibit a range, with the lowest value of 0.947 (corresponding to the NB model) and the highest value of 0.987 (associated with the NBCO

model). Interpreting the R^2, it is apparent that the NBCO model, with the highest R^2, indicates superior model performance. The NBBW model, which achieved a R^2 of 0.975, closely follows as the second-best performer.

RMSE and MSE represent the amount of error according to their definitions. The smaller values of these metrics indicate better model performance. For the NB model, the highest values of RMSE and MSE are 2.050 and 4.204, respectively. In contrast, for the NBCO model, these values are significantly lower at 1.377 and 1.896, demonstrating the superior performance of the NBCO model.

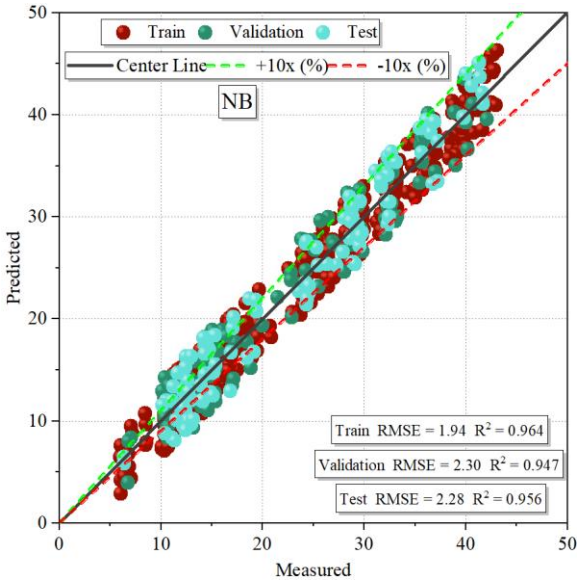
U95, representing data uncertainty, shows that a model's performance improves as this value decreases. According to Table II, the U95 values for models NBCO, NBBW, and NB are 3.747, 4.676, and 5.677, respectively.

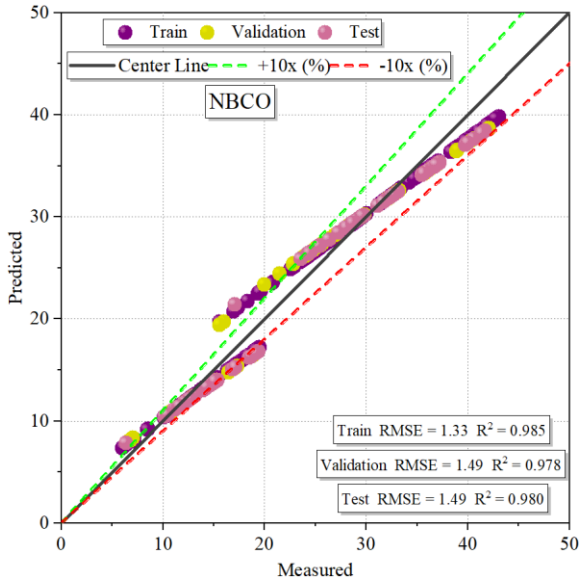
The metric IOA indicates the agreement or accuracy of the model predictions associated to the practical data, and its value falls within the range of 0 to 1. NBCO, with an IOA of 0.995, ranks the highest, demonstrating superior performance. NBBW and NB models are second and third in model quality, respectively.

TABLE II. THE RESULT OF THE DEVELOPED MODELS

Model	*Phase*	*Index values*				
		RMSE	*R2*	*MSE*	*U95*	*IOA*
NB	*Train*	1.940	0.964	3.764	5.375	0.991
	Validation	2.298	0.947	5.282	6.354	0.986
	Test	2.278	0.956	5.188	6.278	0.988
	All	2.050	0.960	4.204	5.677	0.990
NBCO	*Train*	1.325	0.985	1.757	3.598	0.996
	Validation	1.491	0.978	2.222	4.104	0.994
	Test	1.490	0.980	2.219	4.034	0.994
	All	1.377	0.983	1.896	3.747	0.995
NBBW	*Train*	1.605	0.975	2.575	4.429	0.994
	Validation	1.880	0.967	3.536	5.115	0.991
	Test	1.914	0.964	3.662	5.276	0.991
	All	1.698	0.972	2.882	4.676	0.993

Fig. 2. displays a scatter plot for hybrid models, illustrating the variation among predicted and measured values. This scatter plot is generated using RMSE and R^2 values, which primarily influence data dispersion. A decrease in RMSE corresponds to an increase in data density. Furthermore, a higher R^2 value indicates a more precise fit of the line to the data. Based on the visual representations in the plots, it is evident that three primary lines can be identified: a central line, a line representing a 10% overestimation, and a line depicting a 10% underestimation. After explicit consideration, it is apparent that the minimum R^2 value, at 0.947, is associated with model NB, whereas model NBCO exhibits the highest value, 0.985. Furthermore, the highest RMSE is observed in model NB, which is equal to 2.30, and the lowest value for model NBCO is 1.33, representing a 47% reduction in error. Based on these findings, it can be concluded that NBCO is the superior choice for predicting heating load.





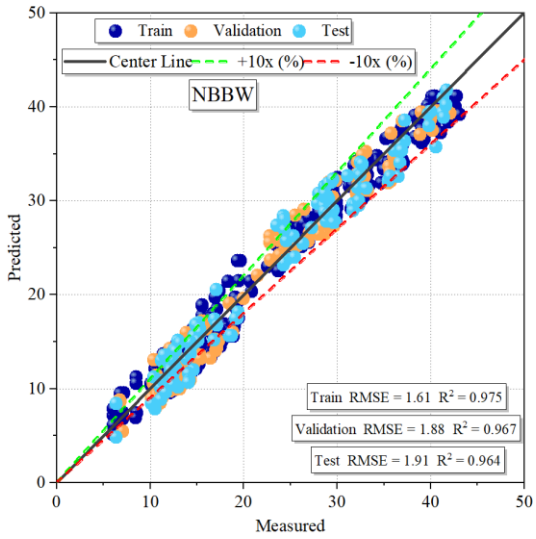


Fig. 2. The hybrid model's created scatter plot.

Fig. 3 illustrates the variations in error metrics ($RMSE$ and MSE) and R^2 values across the three models in this study. According to the trend lines for RMSE and MSE, it is observed that errors in all models initially increase during the train phase. However, there is a noticeable decrease from the validation phase to the test. In summary, after comparing the error rates of RMSE and MSE, it can be deduced that NBCO, with values of 1.377 and 1.896, is the most accurate prediction model, while NBBW and NB are the second and third-ranking models, individually. In all phases, the value of R^2 for NBCO is higher than NBBW and NB by approximately 1.13% and 2.396%, which again shows the superiority of the NBCO.

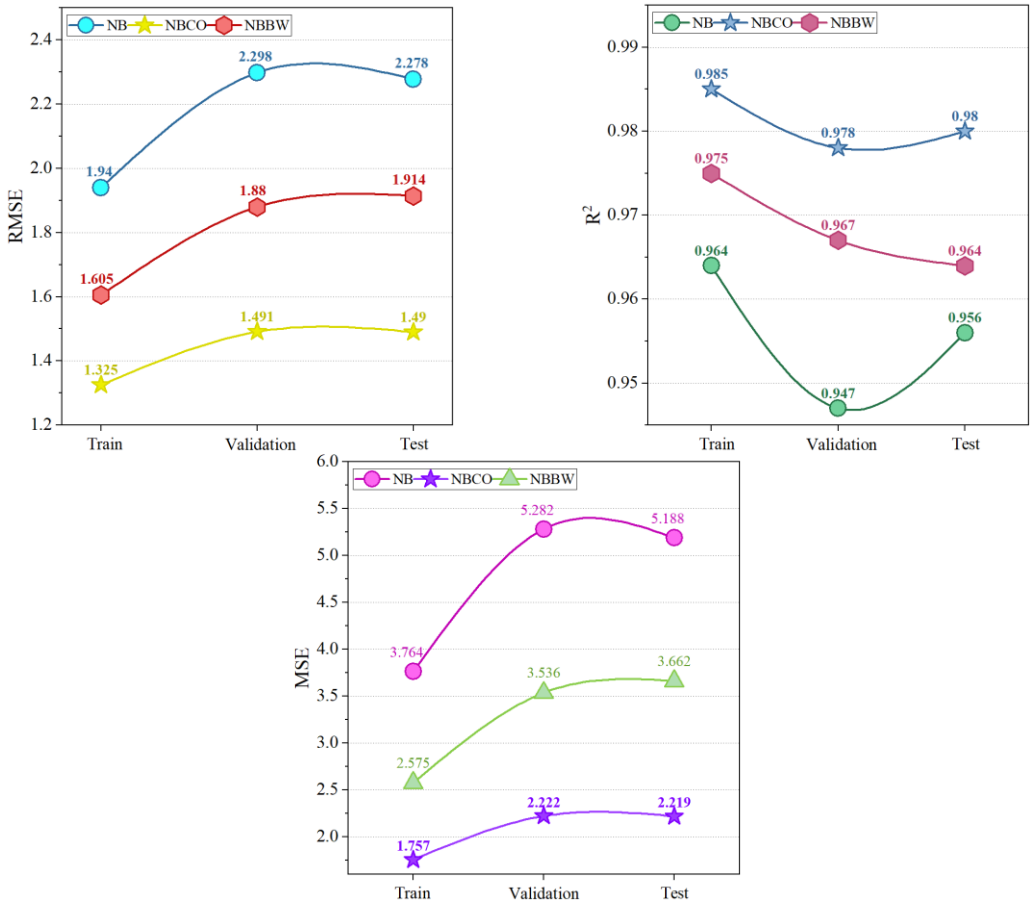


Fig. 3. Comparison between models corresponding RMSE, R^2, MSE.

The histogram distribution diagram, depicting the error percentages of models, is presented in Fig. 4. The horizontal axis displays the percentage of errors, while the vertical axis represents the frequency of occurrences for each model during the training, validation, and test phases. In the basic *NB* model, the error percentage falls within the range of approximately -40 to 40, with the highest frequency around 90. In the case of the two subsequent hybrid models, the error ranges for NBCO and NBBW are approximately -20 to 20 and -30 to 30, correspondingly. The highest frequencies of error values near zero percent for NBCO and NBBW are 100 and 120, respectively.

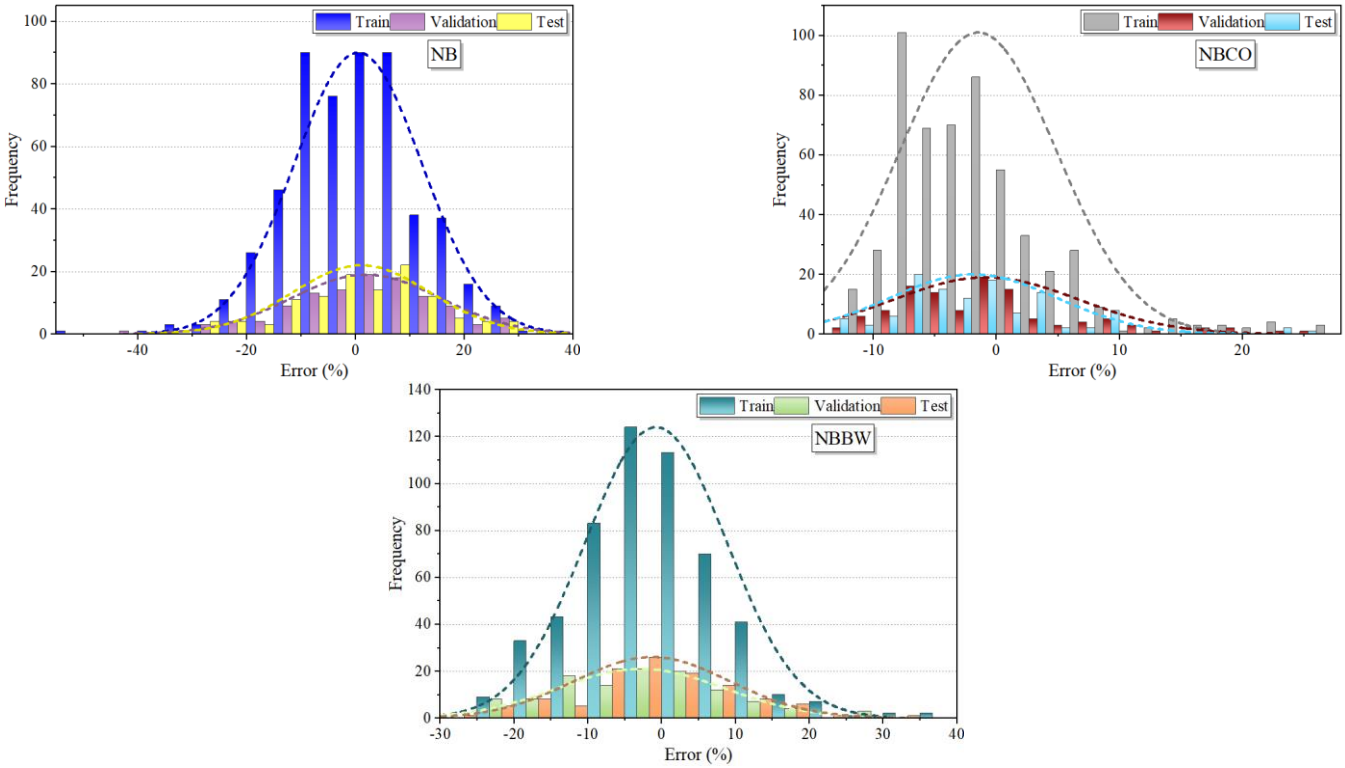


Fig. 4. The error ratio for the hybrid models is created on the Histogram distribution plot.

The multi-line diagram illustrating the error percentages of the models is presented in Fig. 5. The horizontal *axis* represents the number of samples, and the vertical *axis* is divided into three components: the blue axis represents the error rate of model NB. At the same time, the brown and pink axes correspond to models NBCO and NBBW, respectively. It should be noted that the error percentage range for the NB model spans approximately from -40 to 40 during the train, validation, and test phases. In contrast, for the NBCO model, the range extends from -20 to just above 20, while for the NBBW model, it falls within the range of approximately (-30 to 30).

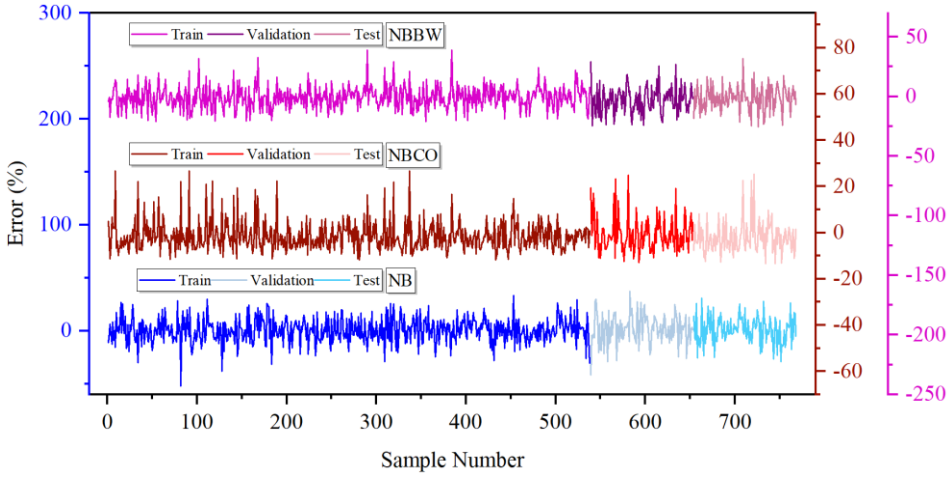


Fig. 5. The Multi-line plot for errors of the developed models.

IV. Discussion

A. Validation of Present Study

The validation of the developed NBCO model in this study demonstrates its superior performance compared to existing models. Table III presents a comparison among the NBCO model from the present study and the PSO-MLP model by Zhou et al. (Zhou u. a. 2020). The NBCO model achieves an R^2 value of 0.985, significantly higher than the 0.9126 obtained by the PSO-MLP model. This indicates that the NBCO model explains a larger proportion of the variance in heating load predictions, showcasing its enhanced predictive power. Additionally, the RMSE of the NBCO model is 1.325, which is substantially lower than the 2.9736 RMSE reported for the PSO-MLP model. A lower RMSE reflects higher accuracy in the predictions, further validating the effectiveness of the NBCO model in accurately forecasting heating loads. These results highlight the advantages of the bio-inspired optimization techniques employed in this study, particularly the Coot Optimization Algorithm, in refining the Naive Bayes model. The validation confirms that the NBCO model outperforms existing approaches, making it a valuable tool for improving energy efficiency and sustainability in building energy management.

TABLE III. The Validation of Developed Model

Article	Model	Evaluator	
		R^2	*RMSE*
Zhou et al. (Zhou u. a. 2020)	PSO-MLP	0.9126	2.9736
Present study	NBCO	0.985	1.325

B. Comparison

Table IV presents a comparative analysis of the best-performing models from the present study alongside similar models from relevant literature, focusing on their ability to predict HL. The models compared include Support Vector Regression (SVR), Multi-Parameter Moving Ridge (MPMR), Light Gradient Boosting Machine (LGBM), and the Naive Bayes optimized with Coot Optimization Algorithm (NBCO) developed in the current study. The SVR model by Moradzadeh et al. (Moradzadeh u. a. 2020) achieved an impressive $RMSE$ of 0.4832 and an R^2 value of 0.9979, indicating a high level of accuracy and predictive power. Similarly, Roy et al. (Roy u. a. 2020) reported an MPMR model with an RMSE of just 0.059 and an R^2 of 0.99, making it one of the most accurate models for heating [39] load prediction. Gong et al. (Gong u. a. 2020) employed the LGBM model, which also performed well, achieving an $RMSE$ of 0.1929 and an R^2 value of 0.9882. In comparison, the NBCO model from the present study produced an $RMSE$ of 1.325 and an R^2 value of 0.985. While the NBCO model's R^2 value is close to those reported in the literature, indicating strong predictive accuracy, its RMSE is notably higher. This suggests that while NBCO captures the overall variance in heating loads [40] effectively, there may be room for improvement in reducing the prediction errors to match or surpass the accuracy levels of the models reported in other studies. Despite this, the NBCO model still offers significant advantages, particularly in its innovative use of bio-inspired optimization techniques. The model's relatively high R^2 value demonstrates its ability to serve as a reliable tool for heating [41] load prediction, with the added potential for further refinement to improve its RMSE. This comparison underscores the value of continuing to explore and optimize machine learning models in the pursuit of enhanced energy efficiency in building management.

TABLE IV. The Comparison of the Best Performed Models Results of Present Study with Some Related Literatures

Articles	Index values			
	Target	*Models*	*RMSE*	R^2
Moradzadeh et al. (Moradzadeh u. a. 2020)	HL	SVR	0.4832	0.9979
Roy et al. (Roy u. a. 2020)	HL	$MPMR$	0.059	0.99
Gong et al. (Gong u. a. 2020)	HL	$LGBM$	0.1929	0.9882
Present Study	HL	$NBCO$	1.325	0.985

C. Limitation

Despite the promising results, this study has several limitations that should be acknowledged, both in the context of the Naive Bayes (NB) model and the broader modeling approach. First, the NB model's inherent assumption of conditional independence among input features may not fully capture the complex interdependencies in building systems, potentially leading to inaccuracies when strong correlations exist between variables such as orientation, glazing area, and thermal performance. This limitation could result in suboptimal predictions, particularly in scenarios where these interactions play a significant role. Additionally, the optimization techniques employed Beluga Whale Optimization (BWO) and Coot Optimization Algorithm (COA) though effective, are relatively novel and less established than traditional methods. Their efficacy in various contexts remains to be thoroughly validated, and there may be cases where these optimizers do not provide substantial improvements over more conventional approaches. Moreover, the study focuses exclusively on predicting heating loads, overlooking other critical aspects of building energy management, such as cooling loads and ventilation. This narrow focus limits the comprehensiveness of the model and its applicability in broader energy efficiency strategies. Finally, the existing model does not account for real-time data integration or adaptive learning, which are increasingly important in dynamic energy management systems. The absence of these features may restrict the model's effectiveness in responding to changing conditions and optimizing performance over time.

V. Conclusion

The contemporary challenge of effectively managing building energy consumption, particularly in structures equipped with air conditioning systems, necessitated a holistic understanding of energy resources and end-uses within buildings. Achieving energy efficiency and sustainability in the built environment demanded the development of optimal predictive tools for estimating building energy consumption. Various modeling methodologies, including traditional approaches based on building geometry and advanced machine

learning models like $ANNs, SVM$, and random forests (RF), were explored for this purpose. Additionally, integrating metaheuristic algorithms emerged as a promising avenue for optimizing these models.

This study extended these efforts by applying the Naive Bayes (NB) model to predict heating loads in buildings and optimize the train process using the Beluga Whale Optimization (BWO) and Coot Optimization Algorithm (COA). Comparative analysis revealed that the optimized NB models outperformed traditional NB, demonstrating the potential for these bio-inspired optimization techniques to enhance predictive models and contribute to greater energy efficiency and sustainability in the built environment. Based on comparative analysis based on numerical values obtained for each evaluation metric corresponding to the developed models, the NBCO hybrid model attained a maximum coefficient of determination of 0.985, surpassed NBBW and NB by 1.03% and 2.2%, respectively, and exhibited minimal performance RMSE error of 1.325, which are notably 17.4% and 31.7% lower than those observed in NBBW and NB. This research served as a significant step toward addressing the energy challenges faced by contemporary facility management, presenting a promising path for future developments in the field.

ACKNOWLEDGMENT

The subject of Artificial Intelligence Industry Application Research Center Facing the Belt and Road Initiative of Zhejiang Business Technology Institute.

REFERENCES

[1] Neto AH, Fiorelli FAS. Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. Energy Build 2008;40:2169–76.

[2] Wei Y, Zhang X, Shi Y, Xia L, Pan S, Wu J, et al. A review of data-driven approaches for prediction and classification of building energy consumption. Renewable and Sustainable Energy Reviews 2018;82:1027–47.

[3] Saffari M, de Gracia A, Ushak S, Cabeza LF. Passive cooling of buildings with phase change materials using whole-building energy simulation tools: A review. Renewable and Sustainable Energy Reviews 2017;80:1239–55.

[4] Dogan T, Reinhart C. Shoeboxer: An algorithm for abstracted rapid multi-zone urban building energy model generation and simulation. Energy Build 2017;140:140–53.

[5] Zhao H, Magoulès F. A review on the prediction of building energy consumption. Renewable and Sustainable Energy Reviews 2012;16:3586–92.

[6] Park JS, Lee SJ, Kim KH, Kwon KW, Jeong J-W. Estimating thermal performance and energy saving potential of residential buildings using utility bills. Energy Build 2016;110:23–30.

[7] Yezioro A, Dong B, Leite F. An applied artificial intelligence approach towards assessing building performance simulation tools. Energy Build 2008;40:612–20.

[8] Yan D, Xia J, Tang W, Song F, Zhang X, Jiang Y. DeST—An integrated building simulation toolkit Part I: Fundamentals. Build Simul, vol. 1, Springer; 2008, p. 95–110.

[9] Crawley DB, Lawrie LK, Winkelmann FC, Buhl WF, Huang YJ, Pedersen CO, et al. EnergyPlus: creating a new-generation building energy simulation program. Energy Build 2001;33:319–31.

[10] York DA, Cappiello CC, Olson KH. DOE-2 Reference Manual: Version 2.1 C. Los Alamos National Laboratory, Solar Energy Group; 1984.

[11] O'Neill Z, O'Neill C. Development of a probabilistic graphical model for predicting building energy performance. Appl Energy 2016;164:650–8.

[12] Yu Z, Haghighat F, Fung BCM, Yoshino H. A decision tree method for building energy demand modeling. Energy Build 2010;42:1637–46. https://doi.org/10.1016/j.enbuild.2010.04.006.

[13] Dimitrov D, Abdo H. Tight independent set neighborhood union condition for fractional critical deleted graphs and ID deleted graphs. Discrete and Continuous Dynamical Systems-S 2019;12:711–21.

[14] Gao W, Guirao JLG, Basavanagoud B, Wu J. Partial multi-dividing ontology learning algorithm. Inf Sci (N Y) 2018;467:35–58.

[15] Catalina T, Virgone J, Blanco E. Development and validation of regression models to predict monthly heating demand for residential buildings. Energy Build 2008;40:1825–32.

[16] behnam Sedaghat, Tejani GG, Kumar S. Predict the Maximum Dry Density of soil based on Individual and Hybrid Methods of Machine Learning. Advances in Engineering and Intelligence Systems 2023;002. https://doi.org/10.22034/aeis.2023.414188.1129.

[17] Masoumi F, Najjar-Ghabel S, Safarzadeh A, Sadaghat B. Automatic calibration of the groundwater simulation model with high parameter dimensionality using sequential uncertainty fitting approach. Water Supply 2020;20:3487–501. https://doi.org/10.2166/ws.2020.241.

[18] Akbarzadeh MR, Ghafourian H, Anvari A, Pourhanasa R, Nehdi ML. Estimating Compressive Strength of Concrete Using Neural Electromagnetic Field Optimization. Materials 2023;16:4200.

[19] Kalogirou SA, Bojic M. Artificial neural networks for the prediction of the energy consumption of a passive solar building. Energy 2000;25:479–91.

[20] Pao H-T. Comparing linear and nonlinear forecasts for Taiwan's electricity consumption. Energy 2006;31:2129–41.

[21] Ben-Nakhi AE, Mahmoud MA. Cooling load prediction for buildings using general regression neural networks. Energy Convers Manag 2004;45:2127–41.

[22] Dong B, Cao C, Lee SE. Applying support vector machines to predict building energy consumption in tropical region. Energy Build 2005;37:545–53.

[23] Nilashi M, Dalvi-Esfahani M, Ibrahim O, Bagherifard K, Mardani A, Zakuan N. A soft computing method for the prediction of energy performance of residential buildings. Measurement 2017;109:268–80.

[24] Gao W, Alsarraf J, Moayedi H, Shahsavar A, Nguyen H. Comprehensive preference learning and feature validity for designing energy-efficient residential buildings using machine learning paradigms. Appl Soft Comput 2019;84:105748.

[25] Li Q, Meng Q, Cai J, Yoshino H, Mochida A. Predicting hourly cooling load in the building: A comparison of support vector machine and different artificial neural networks. Energy Convers Manag 2009;50:90–6.

[26] Zhao J, Liu X. A hybrid method of dynamic cooling and heating load forecasting for office buildings based on artificial intelligence and regression analysis. Energy Build 2018;174:293–308.

[27] Tien Bui D, Moayedi H, Anastasios D, Kok Foong L. Predicting heating and cooling loads in energy-efficient buildings using two hybrid intelligent models. Applied Sciences 2019;9:3543.

[28] Bahiraei M, Heshmatian S, Goodarzi M, Moayedi H. CFD analysis of employing a novel ecofriendly nanofluid in a miniature pin fin heat sink for cooling of electronic components: Effect of different configurations. Advanced Powder Technology 2019;30:2503–16.

[29] Zhong C, Li G, Meng Z. Beluga whale optimization: A novel nature-inspired metaheuristic algorithm. Knowl Based Syst 2022;251:109215. https://doi.org/https://doi.org/10.1016/j.knosys.2022.109215.

[30] Naruei I, Keynia F. A new optimization method based on COOT bird natural life model. Expert Syst Appl 2021;183:115352.

[31] Lowd D, Domingos P. Naive Bayes models for probability estimation. Proceedings of the 22nd international conference on Machine learning, 2005, p. 529–36.

[32] Sibyan H, Svajlenka J, Hermawan H, Faqih N, Arrizqi AN. Thermal Comfort Prediction Accuracy with Machine Learning between Regression Analysis and Naïve Bayes Classifier. Sustainability 2022;14:15663.

[33] Song G, Ai Z, Zhang G, Peng Y, Wang W, Yan Y. Using machine learning algorithms to multidimensional analysis of subjective thermal comfort in a library. Build Environ 2022;212:108790.

[34] Yılmaz D, Tanyer AM, Toker İD. A data-driven energy performance gap prediction model using machine learning. Renewable and Sustainable Energy Reviews 2023;181:113318.

[35] Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. vol. 2. Springer; 2009.

[36] Piryonesi SM, El-Diraby TE. Role of data analytics in infrastructure asset management: Overcoming data size and quality problems. Journal of Transportation Engineering, Part B: Pavements 2020;146:4020022.

[37] Mantegna RN. Fast, accurate algorithm for numerical simulation of Levy stable stochastic processes. Phys Rev E 1994;49:4677.

[38] Zhou G, Moayedi H, Bahiraei M, Lyu Z. 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 2020;254:120082.

[39] Moradzadeh A, Mansour-Saatloo A, Mohammadi-Ivatloo B, Anvari-Moghaddam A. Performance evaluation of two machine learning techniques in heating and cooling loads forecasting of residential buildings. Applied Sciences 2020;10:3829.

[40] Roy SS, Samui P, Nagtode I, Jain H, Shivaramakrishnan V, Mohammadi-Ivatloo B. Forecasting heating and cooling loads of buildings: A comparative performance analysis. J Ambient Intell Humaniz Comput 2020;11:1253–64.

[41] Gong M, Bai Y, Qin J, Wang J, Yang P, Wang S. Gradient boosting machine for predicting return temperature of district heating system: A case study for residential buildings in Tianjin. Journal of Building Engineering 2020;27:100950.