Prediction of Heating and Cooling Load Using Machine Learning Techniques

1st K. Himasree
Electrical Engineering Department
National Institute of Tech. Warangal
Warangal, India
hs21eeb0b27@student.nitw.ac.in

2nd Altaf Q. H. Badar
Electrical Engineering Department
National Institute of Tech. Warangal
Warangal, India
altafbadar@nitw.ac.in

3rd Khai Phuc Nguyen
Electrical and Electronics Engineering
Ho Chi Minh City Univ. of Tech.
Ho Chi Minh City, Vietnam

4th Pradita Octoviandiningrum Hadi
Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia

2024 23rd National Power Systems Conference (NPSC) | 979-8-3315-1956-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/NPSC61626.2024.10987036

***Abstract*—Accurate prediction models for heating and cooling loads are essential for enhancing building energy efficiency, a key component of sustainable development. This study evaluates the effectiveness of three regression methods—Linear Regression, Random Forest Regression, and Bidirectional Encoder Representations from Transformers-based Regression—in predicting these loads. The dataset comprises various building parameters, with heating and cooling loads as target variables. We aim to identify the most accurate and robust predictive model for load forecasting in buildings. Linear Regression was used as the baseline model and is a widely applied method for forecasting. Random Forest Regression is an ensemble learning approach that has been applied to a number of forecasting problems. It averages predictions from multiple decision trees and is, therefore, able to capture non-linear relationships in the most accurate ways. BERT-based Regression, though initially designed for natural language processing tasks, has also been utilized to solve some forecasting problems. Its capability to handle large datasets and complex relationships furthers its case in prediction applications. The results underscore Random Forest Regression as the most effective method for predicting building energy loads. This study utilized Python and its various machine-learning libraries to implement and compare the above methods.**

***Index Terms*—machine learning, energy efficiency, load forecasting, regression methods, building energy management**

I. Introduction

Over the past few decades, the concept of energy-efficient buildings has become increasingly important to reduce energy consumption, minimize energy wastage, and mitigate adverse environmental effects such as CO_2 emissions, etc. [1]. Energy efficiency policies are now crucial in the global energy market, with many countries and institutions prioritizing them to reduce energy consumption significantly. Various global organizations and governments at both national and international levels are focused on improving the energy performance of buildings, including residential and commercial structures.

The three sectors that consume the most energy are industry, transportation, and buildings. According to the International Energy Agency (IEA), residential buildings consume more energy than commercial ones due to their higher numbers. Monitoring and controlling residential consumption is essential due to its significant impact on social welfare, reflecting a substantial demand [2].

Heating and cooling are the most significant energy needs, so their usage should be carefully controlled. A building's heating and cooling load refers to the amount of energy required to maintain a comfortable indoor temperature. Proper building design can significantly reduce energy demand, and enhance energy efficiency [3]. Factors such as design, size, shape, location, and climate influence the heating and cooling load calculations. Installing systems with inappropriate load capacities leads to energy wastage and environmental issues.

Countries in the European Union, United States, and China lead in residential energy consumption, with EU and US each accounting for 40% of their total energy use. European nations, in particular, must comply with legal standards regulating minimum energy efficiency requirements [4].

Possessing comprehensive knowledge of the building's performance and surrounding conditions is essential for managing and optimizing energy use in buildings [5]. The most significant energy resources in a building are electricity, gas, and heating supplies; nevertheless, significant end uses include Heating, Ventilation, and Air Conditioning (HVAC), elevators, and other appliances [6]. In buildings, temperatures are primarily controlled by heating and air conditioning systems [7].

Advances in technology have introduced various building simulation tools that can accurately estimate a building's energy consumption. Machine learning techniques have emerged as an effective solution for predicting energy consumption in residential buildings [8], [9].

In this paper, we use machine learning techniques to predict the Heating and Cooling Load of buildings because the target variables (heating and cooling loads) are continuous values. This approach allows us to accurately model the relationship between building characteristics, environmental factors, and the resulting energy loads. Heating load is the heat needed per unit time to maintain a building's temperature, while cooling load is the heat that must be removed. Reduced energy use

979-8-3315-1956-8/24/$31.00 ©2024 IEEE

lessens demand, cutting fossil fuel consumption and greenhouse gas emissions, thus helping combat global warming [10]. However, the accuracy of these simulations can vary depending on the software and the specifics of the building design. To address this, machine learning tools are increasingly used to analyze the impact of different building parameters. Understanding the complex, non-linear interactions between factors such as building parameters is challenging with traditional methods. Therefore, machine learning models, are suitable for capturing these intricate relationships more effectively. These models can handle high-dimensional data, which is crucial for considering the numerous factors influencing heating and cooling loads. This approach is simple, fast, and provides quantitative analysis, delivering quick results when the model is well-trained and input variables are adjusted.

This study aims to address the critical need for identifying accurate prediction models to enhance the energy efficiency of buildings. The study evaluates the effectiveness of three regression methods—Linear Regression, Random Forest Regression, and Bidirectional Encoder Representations from Transformers-based Regression—in predicting heating and cooling loads based on various building parameters. By leveraging these machine learning techniques, the study seeks to provide a robust predictive framework that can inform better design and operational strategies for prediction of heating and cooling load, ultimately contributing to reduced energy demand and enhanced sustainability in the building sector.

The subsequent sections of this paper are structured as follows: Section II provides an overview of the case study and the dataset utilized. Section III details the methodologies and techniques employed. Section IV presents the findings, and Section V offers the concluding remarks.

II. Case Study

The buildings considered in the dataset differ in glazing area, orientation, etc., amongst other parameters. The dataset contains 768 samples distributed over 8 features for the prediction of heating and cooling loads. The eight attributes in the dataset are denoted by X1…X8, whereas the two responses are denoted by Y1 and Y2 as given in Table I.

TABLE I: Nomenclature of Features

Symbol	VARIABLE	Symbol	VARIABLE
Y1	Heating Load	Y2	Cooling Load
X1	Relative Compactness	X2	Surface Area
X3	Wall Area	X4	Roof Area
X5	Overall Height	X6	Orientation
X7	Glazing Area	X8	Glazing Area Distribution

In Figures 1 and 2, the variables of relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution are depicted on the x-axis, while the heating load and cooling load on the y-axis. A statistical analysis for the considered data is also performed and the statistical description of data is given in Figure 3. Based on the analysis, each attribute in the dataset contains 768 values, confirming the absence of missing or NaN values. Therefore, no data cleaning is required. Additionally, Figure 3 presents the mean, standard deviation, minimum, maximum, and percentiles (25th, 50th, and 75th) for each attribute. Further analysis assess the correlation coefficients between the input and output variables. The correlation matrix depicted in the Figure 4 reveal varying degrees of correlation among the variables. Despite some variables showing high correlation and others lower, none were excluded due to the relatively small number of features involved.

III. Prepare Your Paper Before Styling

A. Evaluation Metric

To evaluate the different methods applied to the dataset and verify their performance, we use the Mean Square Error (MSE), and the formulation is shown in Eq.1.

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

where:

- $\hat{y}_i \rightarrow$ predicted value.
- $y_i \rightarrow$ actual value.
- $n \rightarrow$ number of data points.

B. Linear Regression Method

Linear regression is a foundational technique in machine learning and statistics used for predicting a continuous outcome variable (dependent variable) based on one or more predictor variables (independent variables or features) [11]. Simple Linear Regression equation is shown in Eq.2 below:

$$Y = b_0 + b_1 X + \epsilon \tag{2}$$

Where:

- $X \rightarrow$ independent variable.
- $Y \rightarrow$ dependent variable.
- $b_1 \rightarrow$ slope
- $b_0 \rightarrow$ intercept
- $\epsilon \rightarrow$ error

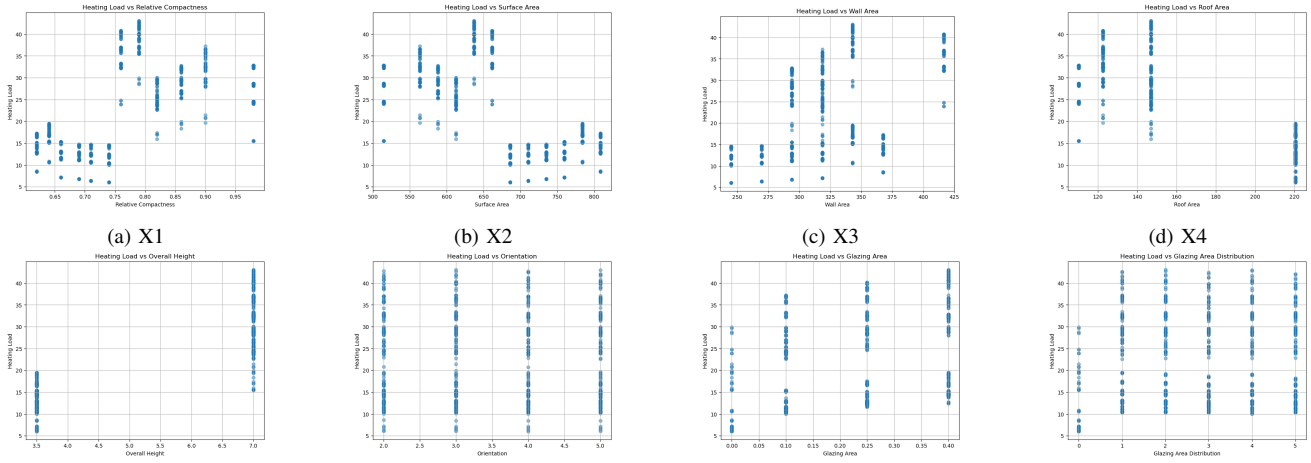
Multiple Linear Regression equation is shown in Eq.3 below:

$$Y = b_0 + b_1 X_1 + \cdots + b_n X_n + \epsilon \tag{3}$$

The goal is to find coefficients $b_0, b_1, \cdots, b_n$ minimizing the sum of squared differences between predicted Y and actual Y in training data. The process involves optimization techniques, often using the least squares method.

Algorithm of Linear Regression for Predicting Heating and Cooling Load:

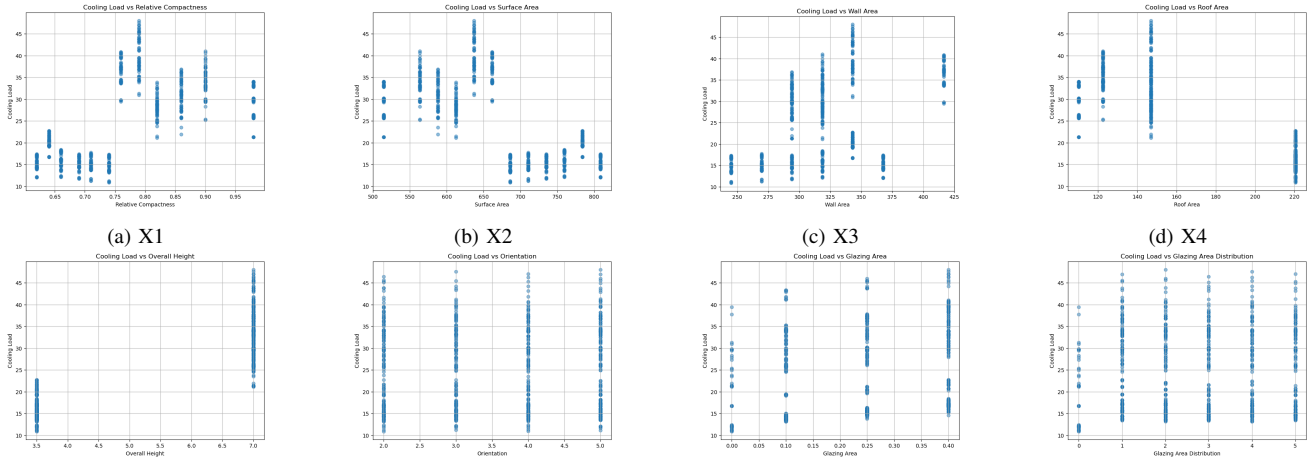
1) **Data Preparation:** Features (X) and target variables (Y) are separated from the dataset.
2) **Training and Testing:** The dataset is split into training and testing sets using a predefined ratio.
3) **Model Training:** Using the training set, a linear regression model is trained to predict the target variable.



(a) X1 (b) X2 (c) X3 (d) X4

(e) X5 (f) X6 (g) X7 (h) X8

Fig. 1: Input Data for predicting Heating Load



(a) X1 (b) X2 (c) X3 (d) X4

(e) X5 (f) X6 (g) X7 (h) X8

Fig. 2: Input Data for Predicting Cooling Load

4) **Model Evaluation:** The model's performance is evaluated using metrics such as variance score and mean squared error (MSE).
5) **Visualization:** Results are visualized, often with scatter plots showing predicted vs. actual values.

C. Random Forests

Random Forest is an ensemble learning method widely used in machine learning for both classification and regression tasks [12]. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees as shown in Eq.4.

$$\hat{y} = \frac{1}{N_{trees}} \sum_{i=1}^{N_{trees}} f_i(X) \tag{4}$$

where:

- $\hat{y}$ is the predicted value,
- $f_i(X)$ is the prediction from the i-th decision tree,
- N_{trees} is the number of trees in the forest.

Algorithm of Random Forests for predicting Heating and Cooling Load:

1) **Data Preparation:** Features (X) and target variables (Y) are prepared similarly as for linear regression.
2) **Training and Testing:** The dataset is split into training and testing sets.
3) **Model Initialization:** Multiple decision trees (forest) are initialized based on random subsets of data.
4) **Model Training:** Each decision tree is trained independently using the training set.
5) **Model Evaluation:** Aggregate predictions from all trees are used to evaluate performance using metrics like MSE.
6) **Visualization:** Predicted vs. actual values are visualized to assess model accuracy.

	X1	X2	X3	X4	X5	X6
count	768.000000	768.000000	768.000000	768.000000	768.00000	768.000000
mean	0.764167	671.708333	318.500000	176.604167	5.25000	3.500000
std	0.105777	88.086116	43.626481	45.165950	1.75114	1.118763
min	0.620000	514.500000	245.000000	110.250000	3.50000	2.000000
25%	0.682500	606.375000	294.000000	140.875000	3.50000	2.750000
50%	0.750000	673.750000	318.500000	183.750000	5.25000	3.500000
75%	0.830000	741.125000	343.000000	220.500000	7.00000	4.250000
max	0.980000	808.500000	416.500000	220.500000	7.00000	5.000000

	X7	X8	Y1	Y2
count	768.000000	768.00000	768.000000	768.000000
mean	0.234375	2.81250	22.307195	24.587760
std	0.133221	1.55096	10.090204	9.513306
min	0.000000	0.00000	6.010000	10.900000
25%	0.100000	1.75000	12.992500	15.620000
50%	0.250000	3.00000	18.950000	22.080000
75%	0.400000	4.00000	31.667500	33.132500
max	0.400000	5.00000	43.100000	48.030000

Fig. 3: Data Description

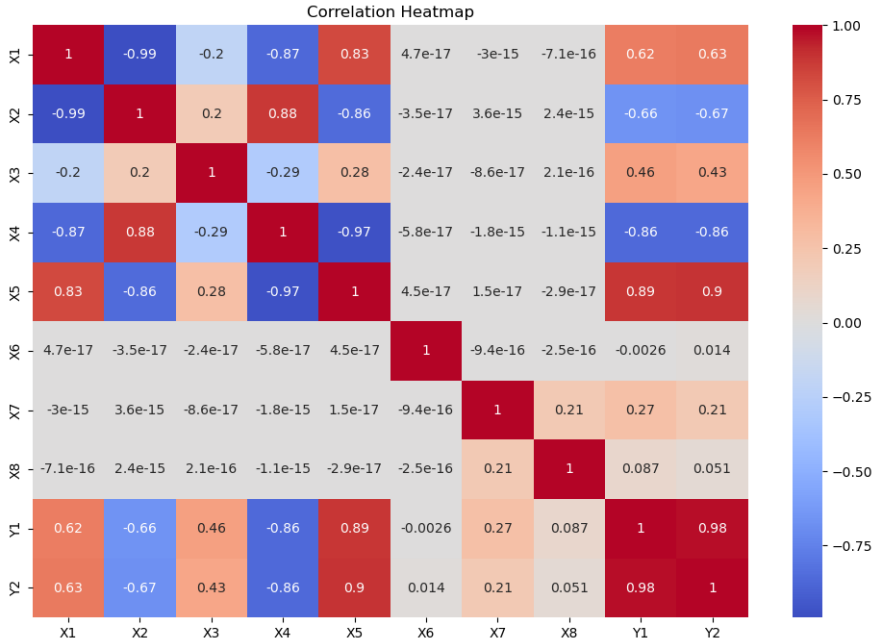


Fig. 4: Correlation between input and output variables

D. Transformers

In machine learning, transformers refer to a type of model architecture widely used for various tasks, especially in Natural Language Processing (NLP) [13]. The term "transformer" can be used in a broader context, referring to both the Transformer model architecture and transformer-based models like Bidirectional Encoder Representations from Transformers (BERT), GPT, and T5.

The following are the tasks for implementation BERT model for regression tasks:

- **Model Input:**
  - Input sequence X is tokenized and prepared as input to BERT.
- **Model Architecture:**
  - BERT processes the tokenized input through multiple transformer layers.
  - It utilizes self-attention mechanisms to capture dependencies between tokens.
- **Regression Head:**
  - A regression-specific head is added to the pre-trained BERT model.
  - The output is a scalar prediction $\hat{y}$.
- **Loss Function:**
  - Mean squared error (MSE) or a similar regression loss function is used to train BERT for regression.
- **Training Objective:**
  - Fine-tuning BERT on a regression-specific dataset to optimize for regression performance.

Algorithm of BERT-based Regression for predicting Heating and Cooling Load:

1) **Data Preparation:** Data is tokenized and converted into tensors suitable for BERT input.
2) **Model Initialization:** Pre-trained BERT models are loaded for sequence classification/regression.
3) **Training Setup:** BERT is fine-tuned on the training set using a regression-specific objective.
4) **Training Iterations:** The model is trained over multiple epochs, adjusting weights based on regression loss.
5) **Model Evaluation:** Predictions are made on the test set, and performance is evaluated using MSE or other regression metrics.
6) **Visualization:** Depending on the task, results may be visualized through scatter plots or regression evaluation plots.

A common algorithm for all three methods applied in this work is given in Algorithm 1. The dataset was divided into 70% and 30% for training and testing of the models, respectively. The number of trees (estimators) utilized in the Random Forest method was 100. In BERT-based regression, the number of epochs was taken as 5, the learning rate was set to $1e^{-5}$, and the batch size was considered to be 32.

IV. Results

The methods discussed in the previous section are applied to the test data considered [3]. The MSE is used to measure the performance of the model, representing the average magnitude of error between the predicted and the original output values. MSE can range between 0 and ∞. A lower MSE value indicates better model performance. The methods are evaluated by comparing the original output values with the predicted output values using the MSE. The MSE values obtained for the different methods applied are displayed in Table II.

TABLE II: Performance Comparison of Regression Methods

Load	Linear Regression	Random Forests	Transformers
Heating Load	10.818	0.281	388.724
Cooling Load	12.393	3.284	472.115

The performance of the three methods was compared based on their ability to predict the target variables $Y1$ and $Y2$. Overall, Random Forest Regression demonstrated the lowest MSE, indicating superior predictive accuracy compared to Linear Regression and BERT-based Regression. Linear Regression showed moderate performance. In contrast, BERT-based Regression, adapted from natural language processing tasks, exhibited competitive performance with potential for further optimization.

Random Forest Regression leverages the power of ensemble learning, combining the predictions of multiple decision trees to achieve better performance. This method was particularly effective in capturing the non-linear relationships in the data, leading to significantly lower MSE values for both heating and cooling loads.

Algorithm 1 Common Algorithm for Linear Regression, Random Forest, and BERT Models

1: **Start**
2: **Read Data**
- Import necessary libraries
- Load the data

3: **Separate Features and Target Variables**
- Separate features (X) from target variables ($Y1$, $Y2$).

4: **Split Data into Training and Testing Sets**
- Split the data into training and testing sets. Set `test_size` to 0.3 and `random_state` to 1 for reproducibility.

5: **Preprocessing for Specific Models**
- Linear Regression and Random Forest: No additional preprocessing needed.
- BERT: Convert the training and testing sets for X and Y to long tensors.

6: **Create DataLoaders (for BERT Model)**
- Create a `TensorDataset` for the training and testing sets of $Y1$ and $Y2$.
- Use `DataLoader` to create loaders for the training and testing datasets of $Y1$ and $Y2$.

7: **Initialize the Model**
- Linear Regression: Initialize `LinearRegression` model.
- Random Forest: Initialize `RandomForestRegressor` model
- BERT: Load the pre-trained BERT model for sequence classification.

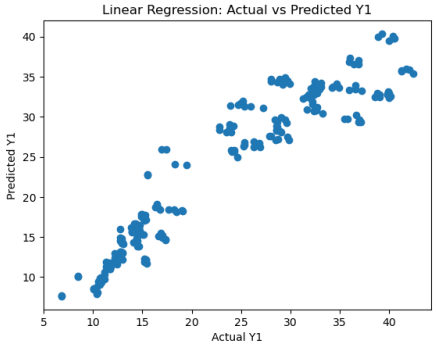
8: **Initialize the Optimizer (for BERT Model)**
- Initialize the optimizer (`AdamW`) for the BERT model parameters with a learning rate of `1e-5`.

9: **Training Loop**
- Train the model using the training data.
- Linear Regression and Random Forest: Fit the model directly using the training data.
- BERT: Perform a training loop with forward pass, backward pass, and optimizer step for a set number of epochs.

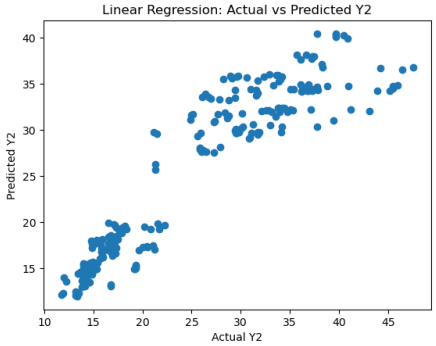
10: **Evaluate the Model**
- Set the model to evaluation mode (if applicable).
- Make predictions on the test set.
- Calculate and print evaluation metric

11: **Visualize Results (if applicable)**
12: **End:Conclude the process.**

Visualizations such as scatter plots or regression evaluation plots were used to validate the predictions against actual values. Figure 5a and 5b, shows the Linear Regression predictions for both loads. Similarly, Figure 6a and 6b, illustrates the predictions by Random Forests. X-axis gives the actual values

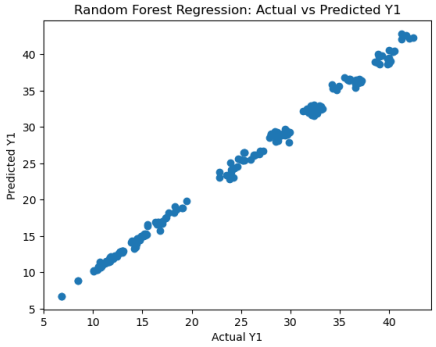


(a) Linear Regression: Actual vs Predicted Y1

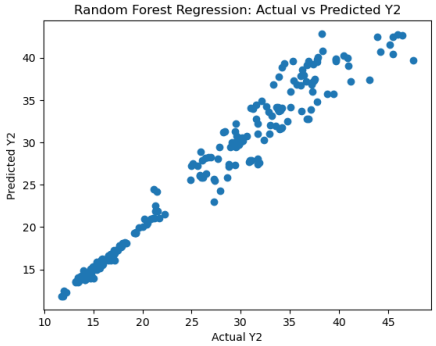


(b) Linear Regression: Actual vs Predicted Y2

Fig. 5: Comparison of Linear Regression Models



(a) Random Forests: Actual vs Predicted Y1



(b) Random Forests: Actual vs Predicted Y2

Fig. 6: Comparison of Random Forest Models

of the load whereas the Y-axis presents the predicted load.

These visualizations highlight the accuracy of the predictions made by the different models. The scatter plots for Linear Regression show a more significant spread of points, indicating less accurate predictions compared to Random Forests. Random Forests' plots show points clustering closer to the line of perfect prediction, demonstrating their superior accuracy.

Conclusion

In this study, we evaluated Linear Regression, Random Forest Regression, and BERT-based Regression for predicting heating and cooling loads in buildings. Performance was assessed using Mean Squared Error (MSE) metrics and visual comparisons. Random Forest Regression achieved the lowest MSE for both heating and cooling loads, proving effective at capturing complex, non-linear relationships. Linear Regression showed moderate performance, providing reasonable predictions but struggling with intricate patterns. BERT-based Regression is not suitable for such prediction.

References

[1] J. Srihari and B. Santhi, "Prediction of heating and cooling load to improve energy efficiency of buildings using machine learning techniques," *J. Mech. Cont. Math. Sci*, vol. 13, no. 5, pp. 97–113, 2018.

[2] A. Moradzadeh, O. Sadeghian, K. Pourhossein, B. Mohammadi-Ivatloo, and A. Anvari-Moghaddam, "Improving residential load disaggregation for sustainable development of energy via principal component analysis," *Sustainability*, vol. 12, no. 8, p. 3158, 2020.

[3] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy and buildings*, vol. 49, pp. 560–567, 2012.

[4] E. Parliament, "Directive 2002/91/ec of the european parliament and of the council of 16 december 2002 on the energy performance of buildings," *Off. J. Eur. Union*, vol. 1, pp. 65–70, 2003.

[5] A. Moradzadeh, A. Mansour-Saatloo, B. Mohammadi-Ivatloo, and A. Anvari-Moghaddam, "Performance evaluation of two machine learning techniques in heating and cooling loads forecasting of residential buildings," *Applied Sciences*, vol. 10, no. 11, p. 3829, 2020.

[6] G. Platt, J. Li, R. Li, G. Poulton, G. James, and J. Wall, "Adaptive hvac zone modeling for sustainable buildings," *Energy and Buildings*, vol. 42, no. 4, pp. 412–421, 2010.

[7] M. W. Ahmad, M. Mourshed, B. Yuce, and Y. Rezgui, "Computational intelligence techniques for hvac systems: A review," in *Building Simulation*, vol. 9. Springer, 2016, pp. 359–398.

[8] A. Moradzadeh, B. Mohammadi-Ivatloo, M. Abapour, A. Anvari-Moghaddam, and S. S. Roy, "Heating and cooling loads forecasting for residential buildings based on hybrid machine learning applications: A comprehensive review and comparative analysis," *Ieee Access*, vol. 10, pp. 2196–2215, 2021.

[9] R. Chaganti, F. Rustam, T. Daghriri, I. d. l. T. Díez, J. L. V. Mazón, C. L. Rodríguez, and I. Ashraf, "Building heating and cooling load prediction using ensemble machine learning model," *Sensors*, vol. 22, no. 19, p. 7692, 2022.

[10] H. Moayedi, D. T. Bui, A. Dounis, Z. Lyu, and L. K. Foong, "Predicting heating load in energy-efficient buildings through machine learning techniques," *Applied Sciences*, vol. 9, no. 20, p. 4338, 2019.

[11] S. Weisberg, *Applied linear regression*. John Wiley & Sons, 2005, vol. 528.

[12] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.