2026 4th International Conference on Integrated Circuits and Communication Systems (ICICACS) | 979-8-3315-8019-3/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICICACS68679.2026.11578650

AI-DRIVEN PREDICTION OF BUILDING ENERGY AND COMFORT

Darwin R Solomon
Data Science and Business Systems
SRM Institute of Science and Technology
Kattankulathur, Chennai, Tamil Nadu 603203
ds0576@srmist.edu.in

Harsh Kumar
Data Science and Business Systems
SRM Institute of Science and Technology
Kattankulathur, Chennai, Tamil Nadu 603203
hk4519@srmist.edu.in

A P Mohammed Salman
Data Science and Business Systems
SRM Institute of Science and Technology
Kattankulathur, Chennai, Tamil Nadu 603203
as5043@srmist.edu.in

Jeba Sonia J
Associate Professor, Data Science and Business Systems
SRM Institute of Science and Technology
Kattankulathur, Chennai, Tamil Nadu 603203
jebas@srmist.edu.in

***Abstract*—**Since the building sector accounts for around 40% of global energy consumption and 33% of greenhouse gas emissions, energy-efficient building design is an essential objective for sustainable development. This paper presents the Climate AI Design Predictor, an end-to-end, machine-learning-powered framework that uses climate conditions and architectural design parameters to forecast building energy consumption (heating load, cooling load, and total energy consumption) and the thermal comfort index. Meteorological data from five Indian cities—Bangalore, Chennai, Delhi, Jaipur, and Shimla—obtained from NASA's POWER API is combined with the UCI Energy Efficiency dataset. Two ensemble learning algorithms, Random Forest (RF) and Light Gradient Boosting Machine (LightGBM), are trained and compared. Experimental results demonstrate exceptional predictive performance, with the best models achieving R2 scores of 0.9972, 0.9680, 0.9926, and 0.9998.

***Index Terms*—**Thermal Comfort, Random Forest, LightGBM, Machine Learning, Climate-Responsive Design, Sustainable Architecture, and Feature Engineering

I. Introduction

The building industry's need for energy has increased due to the fast urbanization of developing countries, especially India [2]. Heating, ventilation, and air conditioning (HVAC) systems account for the largest portion of buildings' nearly 40% global final energy consumption, according to the International Energy Agency (IEA) [1]. For architects and designers who require quick feedback during the conceptual design phase, this poses a serious obstacle. Traditional physics-based simulation tools such as EnergyPlus [3] require detailed inputs and extensive computation time. A convincing substitute is provided by machine learning (ML) techniques, which, after being trained on representative datasets, produce predictions that are almost instantaneous [4], [7].

This paper presents the *Climate AI Design Predictor*, an integrated ML framework that addresses three key challenges:

1) Multi-target Prediction: predicting the thermal comfort index, total energy consumption, heating load, and cooling load all at once using a single feature set that combines architectural and climatic factors.
2) Climate Contextualization: combining city-specific meteorological data from five Indian cities with varying climates—Bangalore, Chennai, Delhi, Jaipur, and Shimla—that fall into the hot-humid, hot-dry, composite, and cold climate zones.
3) Actionable Deployment: delivering a web application with real-time forecasts and rule-based AI design suggestions that is ready for production.

The rest of the paper is structured as follows: Section II examines relevant literature. The system architecture, data sources, preprocessing pipeline, feature engineering, and machine learning algorithms are all covered in Section III. Section IV presents experimental results. Section V discusses findings and implications. Section VI concludes the paper with future directions.

II. Related Work

The application of machine learning techniques for building energy prediction has garnered significant attention in academic research over the past decade. This section provides a comprehensive review of existing literature, organized into four thematic areas: (i) machine learning for building energy prediction, (ii) ensemble methods and boosting algorithms, (iii) thermal comfort modeling and prediction, and (iv) climate-responsive design and feature engineering approaches.

A. Machine Learning for Building Energy Prediction

Tsanas and Xifara [5] created the foundation setting for building energy prediction research with the UCI Energy Efficiency dataset. Their work showed that eight architectural parameters were employed in iteratively honing regression models to estimate heating and cooling loads in such a way that R^2 values exceeded 0.97. This seminal work sparked an enormous amount of research on data-driven methodologies in the area of building energy modeling, highlighting the potential for machine learning to replace computationally intensive physics-based simulations.

B. Ensemble Methods and Boosting Algorithms

Ensemble methods for predicting building energy have attracted considerable interest in recent years, with studies regularly showing that they outperform single-model methods [9]. Given Breiman's Random Forest algorithm [13], its robustness to overfitting and ability to handle high-dimensional feature spaces have been found to be most useful in building energy applications. Ahmad et al. [6] performed a broad comparison between random forests and artificial neural networks for prediction of high-resolution building energy consumption.

C. Thermal Comfort Modeling and Prediction

The prediction of thermal comfort in buildings has usually used the Predicted Mean Vote (PMV) model that was proposed by Fanger [10]. Based on six factors namely air temperature, mean radiant temperature, relative air velocity, humidity, clothing insulation, and metabolic rate. Thom [11] proposed a simpler Discomfort Index as an alternative thermal assessment metric. As a theoretical comfort prediction module, the PMV model is relatively stable in theory, but it lacks robustness with measuring mean radiant temperature and metabolic rate in real-time in building operations.

D. Climate-Responsive Design and Feature Engineering

This is a key move in the development and implementation of climate-responsive design tools through integrating climate data and building energy prediction. Haidar et al. [16] evaluated data collection period and sensor selection methodologies for smart building occupancy prediction, showing that optimal sensor configurations kept prediction accuracy above 90% while reducing data collection frequency from 1 minute to 20 minutes.

E. Summary of Research Gaps

Although the recent literature on energy prediction via machine learning for buildings is substantial, there are a number of gaps that motivate this study:

1) The majority of existing studies either target energy or thermal comfort prediction, and integrated approaches with few frameworks that simultaneously predict multiple targets from a unified feature set have been few.
2) Comprehensive feature importance assessment across multiple prediction targets and climate contexts is lacking, restricting the generation of actionable design insights.

The thermal comfort index is calculated using the Discomfort Index (DI) formulation shown in Eq. (1).

$$DI = T - 0.55 \times (1 - 0.01 \times RH) \times (T - 14.5) \quad (1)$$

where T is the air temperature (°C) and RH is the relative humidity (%). This equation estimates human thermal discomfort based on temperature and relative humidity conditions.

Earlier work had tackled either energy prediction or thermal comfort as two key goals alone; this study combines both in one platform and has developed it as a full-stack web app featuring climate-contextualized predictions for multiple Indian cities.

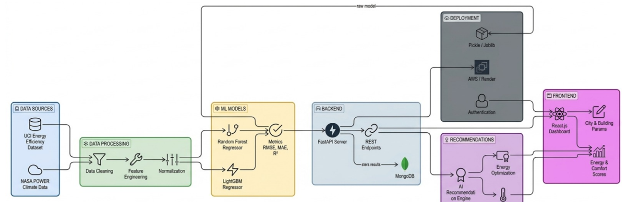

Fig. 1: System architecture of the Climate AI Design Predictor showing the data flow from raw datasets through the ML pipeline to the web application.

TABLE I: Architecture Dataset Feature Mapping

Original	Feature Name	Unit
X_1	Relative Compactness	—
X_2	Surface Area	m^2
X_3	Wall Area	m^2
X_4	Roof Area	m^2
X_5	Overall Height	m
X_6	Orientation	2–5
X_7	Glazing Area	ratio
X_8	Glazing Area Distribution	0–5
Y_1	Heating Load	kWh/m^2
Y_2	Cooling Load	kWh/m^2

III. Methodology

A. System Architecture

Climate AI Design Predictor is designed around four layers with a modular architecture, i.e., (i) data layer, (ii) ML pipeline, (iii) backend API, and (iv) frontend interface (Fig. 1).

B. Data Sources

Two primary datasets are utilized:

1) Climate Dataset

Daily meteorological data is sourced from NASA's Prediction of Worldwide Energy Resources (POWER) API [12] for five Indian cities representing diverse climate zones:

- Bangalore — Tropical savanna (Köppen: *Aw*)
- Chennai — Tropical wet and dry, hot-humid coast
- Delhi — Composite climate with extreme seasonal variation
- Jaipur — Hot and dry semi-arid (*BSh*)
- Shimla — Subtropical highland, cold winters

Each city dataset contains 3,723 daily records (approximately 10 years) with surface air temperature (T_{2M}), relative humidity (RH_{2M}), wind speed (WS_{10M}), and all-sky surface shortwave downward irradiance ($ALLSKY_SFC_SW_DWN$).

2) Architecture Dataset

The UCI Energy Efficiency dataset [5] contains 768 samples with eight architectural building parameters (X_1–X_8) and two target variables (Y_1: heating load, Y_2: cooling load). The original feature names are mapped as shown in Table I.

C. Data Preprocessing Pipeline

Both datasets are processed using a 15-step linear preprocessing pipeline written in Python.

1) Climate Data Preprocessing

The climate preprocessing pipeline involves:

TABLE II: Climate Data Preprocessing Summary

City	Raw	Clean	Outliers	Retained
Bangalore	3,723	3,004	719	80.7%
Chennai	3,723	3,601	122	96.7%
Delhi	3,723	2,766	957	74.3%
Jaipur	3,723	2,853	870	76.6%
Shimla	3,723	3,097	626	83.2%
Total	**18,615**	**15,321**	**3,294**	**82.3%**

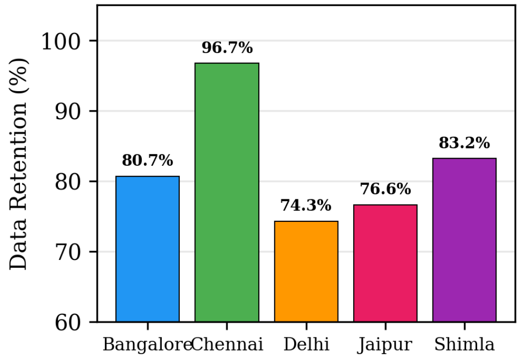

Fig. 2: Data retention percentage per city after preprocessing.

- Data Loading: Auto-discovery and parsing of NASA POWER CSV files, skipping metadata headers.
- Column Formatting: Standardization of column names to lowercase with underscores.
- Sentinel Value Replacement: NASA's missing data indicator (-999) is replaced with NaN. A total of 76 sentinel values per city were identified and replaced.
- Duplicate Removal: No duplicates were detected across the five datasets.
- Data Type Correction: Numeric coercion using `pd.to_numeric`.
- Missing Value Imputation: Median imputation for numeric features; mode imputation for categorical features.
- Invalid Value Filtering: Physically unrealistic readings are removed using domain-specific thresholds as specified in (2):

$$-50 \leq T_{2M} \leq 60; \quad 0 \leq RH_{2M} \leq 100 \tag{2}$$

- Constant Column Removal: Columns with zero variance are dropped.
- Outlier Removal: The Interquartile Range (IQR) method is applied as in (3):

$$x_{\text{valid}} \in [Q_1 - 1.5 \cdot IQR, \quad Q_3 + 1.5 \cdot IQR] \tag{3}$$

- Normalization: Standard scaling is performed via z-score normalization (4):

$$z = \frac{x - \mu}{\sigma} \tag{4}$$

Table II summarizes the preprocessing outcomes for each city. Fig. 2 illustrates the data retention rates across all five cities after preprocessing.

2) Architecture Data Preprocessing

The architecture dataset (768 samples) then goes through the steps of column renaming, data type correction, invalid value filtering, and standard scaling. No duplicates or outliers were removed to yield a 100

D. Feature Engineering

Domain-specific features are engineered to capture non-linear relationships:

1) Climate Features

- Heat Index: A simplified Steadman approximation [14], given by (5):

$$\begin{aligned} HI = &-8.7847 + 1.6114T + 2.3385RH \\ &- 0.1461 \cdot T \cdot RH - 0.0123T^2 - 0.0164RH^2 \\ &+ 0.0022T^2RH + 0.0007T \cdot RH^2 \\ &- 3.6 \times 10^{-6}T^2RH^2 \end{aligned} \tag{5}$$

- Cooling Degree Days: $CDD = \max(0, T_{2M} - 24)$
- Heating Degree Days: $HDD = \max(0, 18 - T_{2M})$
- Temporal Features: Day of year, month, and season (winter, summer, monsoon, post-monsoon based on the Indian climate calendar)

2) Architecture Features

Five interaction features are derived, as defined in (6)–(10):

$$\text{Wall Ratio} = \frac{A_{\text{wall}}}{A_{\text{surface}}} \tag{6}$$

$$\text{Roof Ratio} = \frac{A_{\text{roof}}}{A_{\text{surface}}} \tag{7}$$

$$\text{Glazing} \times \text{Surface} = A_{\text{glazing}} \times A_{\text{surface}} \tag{8}$$

$$\text{Compactness} \times \text{Surface} = RC \times A_{\text{surface}} \tag{9}$$

$$\text{Height} \times \text{Surface} = H \times A_{\text{surface}} \tag{10}$$

E. Dataset Merging

Data points for the climate and architecture are horizontally merged. Climate data (15,321 rows from 5 cities stacked vertically) will be put up alongside architecture data (768 rows). This generates one combined dataset where:

- Energy models use 768 rows with 24 features (11 climate + 13 architecture).
- Comfort models use 15,321 rows with 11 climate-only features.

F. Machine Learning Models

Two state-of-the-art ensemble regression algorithms are employed:

1) Random Forest Regressor (RF)

Random Forest [13] sets up an ensemble of B decision trees, each trained on a bootstrap sample. The final prediction is the average of individual tree predictions, as shown in (11):

$$\hat{y}_{RF} = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{x}) \tag{11}$$

where $T_b(\mathbf{x})$ is the prediction of the b-th tree. The hyperparameters used are: $B = 200$ estimators, max depth $= 15$, min samples split $= 5$, min samples leaf $= 2$, and random state $= 42$.

TABLE III: Model Evaluation Results (Test Set, 20% Split)

Target	**Model**	**MAE**	**RMSE**	$\mathbf{R^2}$
Heating Load	RF	0.3729	0.5355	**0.9972**
	LightGBM	0.4470	0.7311	0.9949
Cooling Load	RF	1.2760	2.0433	0.9549
	LightGBM	1.0917	1.7228	**0.9680**
Energy Consump.	RF	1.1687	1.8084	0.9916
	LightGBM	1.0819	1.6920	**0.9926**
Thermal Comfort	RF	**0.0028**	**0.0059**	**0.9998**
	LightGBM	0.0044	0.0074	0.9998

2) Light Gradient Boosting Machine (LightGBM)

. LightGBM [8] is a gradient boosting framework that grows trees leaf-wise (best-first) rather than level-wise, selecting the leaf with the maximum delta loss for splitting.

The additive model at iteration t is given by (12):

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \eta \cdot h_t(\mathbf{x}) \quad (12)$$

where $\eta = 0.05$ is the learning rate and h_t is the tree learned at iteration t. Additional hyperparameters: $T = 300$ iterations, max depth $= 10$, num leaves $= 31$, min child samples $= 10$.

G. Training Procedure

The dataset is divided into 80/20 groups using a stratified random split with a fixed random seed of 42. It can therefore be replicated, with ease and accuracy. For each of four targets—heating load, cooling load, total energy consumption (= heating + cooling), and thermal comfort (Eq. 1)—both RF and LightGBM are trained separately to obtain eight models altogether.

H. Evaluation Metrics

Three standard regression metrics are used, defined in (13)–(15):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (13)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (14)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (15)$$

where y_i and $\hat{y}_i$ are the actual and predicted values respectively, $\bar{y}$ is the mean of the actual values, and n is the number of test samples. The choice of the best model for each target is based on the highest R^2 score.

IV. Results

A. Model Performance Comparison

Table III presents the complete evaluation results for all eight trained models across the four prediction targets. Table IV summarizes the best-performing model selected for each target based on R^2 score.

Key observations from the results:

- Random Forest gets high predictive power for heating load ($R^2 = 0.9972$, MAE $= 0.37$ kWh/m^2) and thermal comfort ($R^2 = 0.9998$, MAE $= 0.003$), suggesting bagging-based ensembles can capture the deterministic relationships in these targets efficiently.
- LightGBM outperforms RF on cooling load ($R^2 = 0.968$ vs. 0.955) and total energy consumption ($R^2 = 0.993$ vs. 0.992), validating the utility of sequential boosting for targets with more complicated and non-linear relationships.
- All eight models achieve $R^2 > 0.95$, which confirms the validity of the feature engineering and preprocessing pipeline.

Fig. 3, Fig. 4, and Fig. 5 visualize the R^2, MAE, and RMSE comparisons across all targets. Fig. 6 presents the heatmap of R^2 scores across all models and prediction targets.

B. Feature Importance Analysis

Feature importance analysis provides critical insights into which input parameters most influence each prediction target. Fig. 7, Fig. 8, and Fig. 9 show the top-10 feature importances for the best models. Lundberg and Lee [15] proposed SHAP

TABLE IV: Best Model Selection Summary

Target	**Best Model**	**MAE**	**RMSE**	$\mathbf{R^2}$
Heating Load	Random Forest	0.3729	0.5355	0.9972
Cooling Load	LightGBM	1.0917	1.7228	0.9680
Energy Consump.	LightGBM	1.0819	1.6920	0.9926
Thermal Comfort	Random Forest	0.0028	0.0059	0.9998

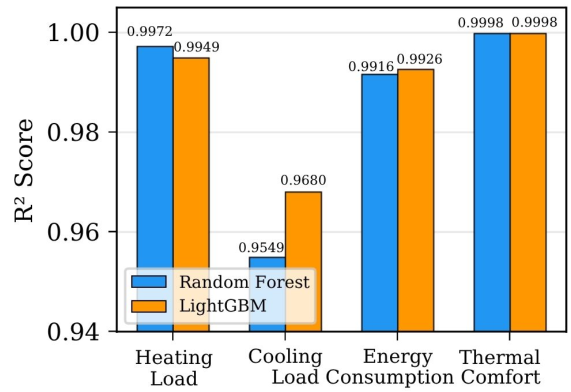

Fig. 3: R^2 score comparison between Random Forest and LightGBM across all four prediction targets. Both models achieve $R^2 > 0.95$ for all targets.

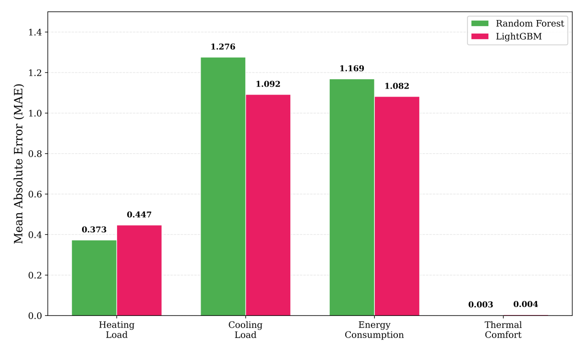

Fig. 4: Mean Absolute Error (MAE) comparison. LightGBM achieves lower MAE for cooling load and energy consumption; RF excels at heating load and thermal comfort.

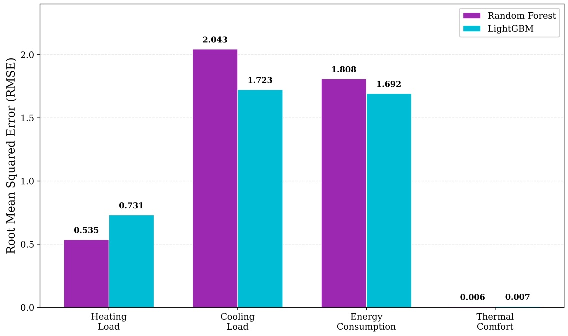

Fig. 5: Root Mean Squared Error (RMSE) comparison across all targets and models.

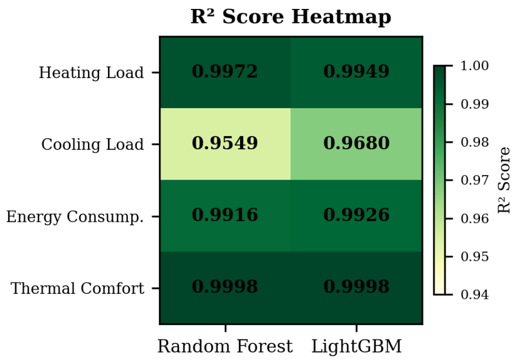

Fig. 6: Heatmap of R^2 scores across models and targets. Darker shades indicate higher performance.

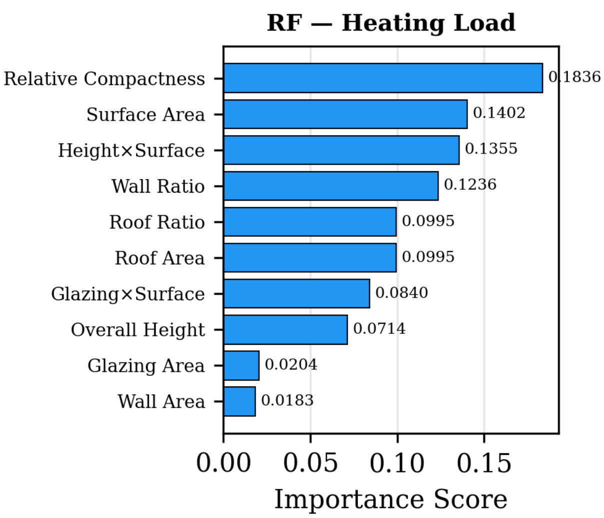

Fig. 7: Feature importance for heating load prediction (RF). Relative compactness, surface area, and height–surface interaction are the top three contributors.

(SHapley Additive exPlanations) as a unified approach for interpreting model predictions, which complements traditional feature importance methods.

Key insights from feature importance:

1) Heating Load is largely determined by architectural parameters (relative compactness: 18.4 and surface area: 14.0), implying that the building form is the most

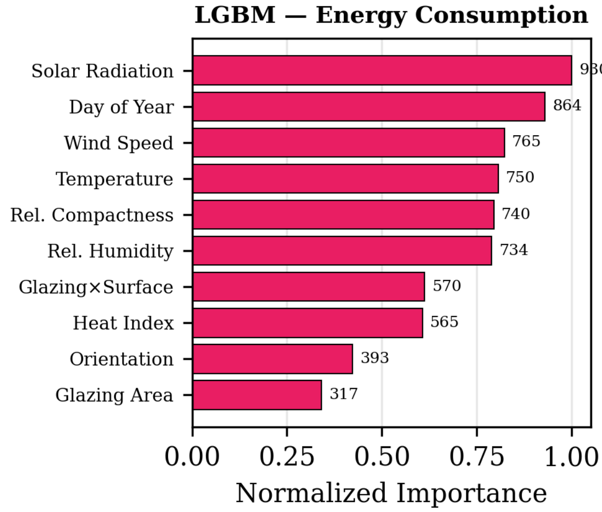

Fig. 8: Feature importance for total energy consumption prediction (LightGBM). Both climate and architectural interaction features contribute significantly.

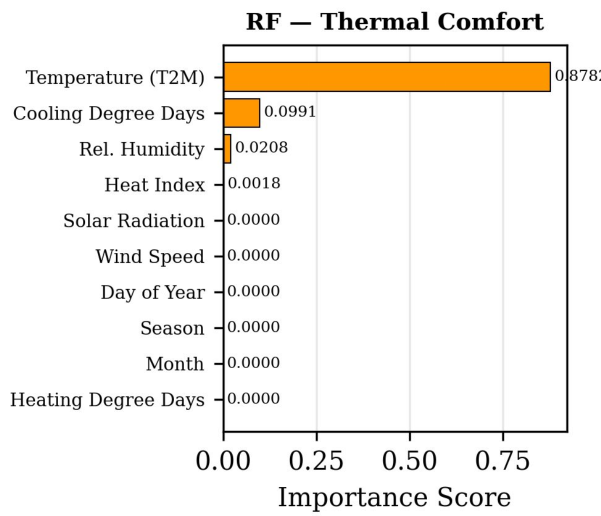

Fig. 9: Feature importance for thermal comfort prediction (RF). Temperature is the dominant contributor at 87.8%, consistent with the Discomfort Index formulation.

important factor in heating needs.

2) Cooling Load and Energy Consumption are strongly correlated to climate parameters
3) Thermal Comfort is almost entirely determined by temperature (87.8%), consistent with the theoretical Discomfort Index formulation (Eq. 1).

V. Discussion

A. Model Selection Trade-offs

The results indicated that RF is complementary to LightGBM: RF works well when the targets have strong links to the input features (heating load, thermal comfort) and LightGBM gets to know the interactions that are both more complex and non-linear (cooling load, energy consumption).

B. Climate-Responsive Insights

Feature importance analysis shows that distinct targets need different groups of influential features Architectural determined targets: Heating loads are mainly determined by building geometry (compactness, surface area, height), indicating that passive design strategies produce maximum effect on heating efficiency.

C. Preprocessing Impact

Data extraction of the preprocessing pipeline removes 3,294 (17.7) climate records from the five climate datasets, with Delhi producing the highest outlier rate (25.7) and Chennai contributing the lowest (3.3).

D. Limitations

Several limitations should be acknowledged:

1) The architecture dataset consists of 768 samples from a single simulation study. A larger, more diverse dataset would improve model generalizability.
2) The horizontal merging method produces a sparse combined dataset. Future work could explore city-building paired simulations.
3) The thermal comfort model employs a simplified Discomfort Index instead of the complete six-parameter Fanger PMV model.
4) Climate data is based on historical averages; real-time weather integration would provide more dynamic predictions.

VI. Conclusion and Future Work

In this article, we introduced the Climate AI Design Predictor, an all in one ML environment for building energy consumption and thermal comfort prediction based on ML. The key contributions are: A multi-target prediction system for all four targets at a $R^2 > 0.96$ threshold with ensemble learning algorithms. This study aimed to integrate climate data from five diverse Indian cities through a predictive model, using predictive tools to make climate-contextualized forecasts.FastAPI + React.js + MongoDB — a full-stack deployment with real-time predictions and AI-generated design recommendations. The most successful models — RF for heating load ($R^2 = 0.9972$), thermal comfort ($R^2 = 0.9998$), LightGBM for cooling load ($R^2 = 0.9680$) and energy consumption ($R^2 = 0.9926$) — show that ensemble ML approaches can quickly and accurately mimic detailed building energy simulation. Future work will focus on:

- Incorporating deep learning models (e.g., neural networks, attention-based architectures) for potentially capturing higher-order interactions.
- Adding real-time weather API integration for dynamic, location-aware predictions.
- Expanding to additional cities and international climate zones.
- Implementing the full PMV-PPD thermal comfort model for six-parameter comfort assessment.
- Deploying the application to a cloud platform (AWS/GCP) with horizontal scaling capabilities.

Acknowledgment

NASA's POWER (Prediction Of Worldwide Energy Resources) project provides the climate data for this study. The architecture dataset is provided by the UCI Machine Learning Repository, contributed by Tsanas and Xifara.

References

[1] International Energy Agency, "Global Status Report for Buildings and Construction 2019," United Nations Environment Programme, 2019.
[2] The Energy and Resources Institute (TERI), "Energy Efficiency in Indian Buildings: Trends and Projections," TERI Policy Brief, 2020.
[3] D. B. Crawley, L. K. Lawrie, F. C. Winkelmann et al., "EnergyPlus: Creating a new-generation building energy simulation program," *Energy and Buildings*, vol. 33, no. 4, pp. 319–331, 2001.
[4] K. Amasyali and N. M. El-Gohary, "A review of data-driven building energy consumption prediction studies," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1192–1205, 2018.
[5] 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.
[6] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption," *Energy and Buildings*, vol. 147, pp. 77–89, 2017.
[7] S. Seyedzadeh, F. P. Rahimian, I. Glesk, and M. L. Roper, "Machine learning for estimation of building energy consumption and performance: a review," *Visualization in Engineering*, vol. 6, no. 1, p. 5, 2018.
[8] G. Ke, Q. Meng, T. Finley et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 3149–3157, 2017.
[9] Z. Wang, T. Hong, and M. A. Piette, "Building thermal load prediction through shallow machine learning and deep learning," *Applied Energy*, vol. 263, p. 114683, 2020.
[10] P. O. Fanger, *Thermal Comfort: Analysis and Applications in Environmental Engineering*. Copenhagen: Danish Technical Press, 1970.
[11] E. C. Thom, "The Discomfort Index," *Weatherwise*, vol. 12, no. 2, pp. 57–61, 1959.
[12] NASA Langley Research Center, "POWER Data Access Viewer," [Online]. Available: https://power.larc.nasa.gov/. Accessed: Mar. 2025.
[13] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[14] R. G. Steadman, "The Assessment of Sultriness. Part I: A Temperature-Humidity Index Based on Human Physiology and Clothing Science," *Journal of Applied Meteorology*, vol. 18, no. 7, pp. 861–873, 1979.
[15] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.
[16] N. Haidar, R. Bouzidi, and M. Oussalah, "Optimizing Data Collection and Sensor Selection for Smart Building Occupancy Prediction," *Energy and Buildings*, vol. 305, p. 113892, 2024.