Data driven techniques for predicting energy efficiency parameters

Giridhar Shanmugam
Undergraduate Student
Vellore Institute of technology
Chennai, India
harikishore.mg2021@vitstudent.ac.in

Harikishore Manday
Undergraduate Student
Vellore Institute of technology
Chennai, India
giridhar.shanmugam2021@vit.ac.in

Nikhil
Undergraduate Student
Vellore Institute of technology
Chennai, India
nikhil.2021e@vitstudent.ac.in

***Abstract*—Abstract—Accurate prediction of energy efficient parameters is essential for improving energy consumption, reducing impacts on the environment and cost cutting.Existing traditional models often face challenges in complexity, non linearity and uncertainity of energy systems.This paper implements different data driven approaches and ensemble techniques including linear regression models, ridge, lasso, random forest, XG Boost and neural networks for predicting energy efficiency in solar panels, thermal power plants etc. by analyzing their building parameters.Performance of all the models was evaluated using Root Mean Square Error(RMSE), R-squared score and cross validation techniques to ensure robustness of the obtained results.The comparitive analysis provides understanding of strengths and weaknesses of each model regarding prediction of accurate parameters of building for efficient energy usage.Among all the models trained, XGBoost exhibited better performance with and RMSE of 0.3926 and R-square score of 0.9985.**

***Index Terms*—Energy,efficiency,parameters,models.**

I. Introduction

with the rise in global energy demands, the need for sustainable and efficient energy management is greater. Traditional methods of assessing energy efficiency depend on empirical models and expert-driven approaches. While these methods have been useful, they struggle to keep up with the real-world conditions that are constantly evolving and changing. In response, data-driven techniques—powered by artificial intelligence (AI), machine learning (ML), and statistical modeling are transforming the way we predict and optimize energy efficiency.

These modern approaches rely on historical data,realtime sensor readings, and environmental factors to anticipate energy consumption patterns, pinpoint inefficiencies, and recommend the smarter operational strategies. Techniques such as regression models, time-series analysis, deep learning, and hybrid optimization frameworks play a crucial role in this transformation. For instance, multiple linear regression (MLR) and support vector regression (SVR) help us to understand how factors like temperature, humidity, and occupancy impacts the energy usage. Meanwhile, the time-series models like ARIMA and Long Short-Term Memory networks (LSTMs) allow us to forecast future energy needs that are based on past trends.

Deep learning, particularly artificial neural networks (ANNs) and convolutional neural networks (CNNs), enhances predictive accuracy by capturing complex relationships in large datasets. Reinforcement learning (RL), where intelligent systems continuously learn and adapt to their energy management strategies through real-time feedbacks and past trends. Hybrid models that blend AI-driven techniques with traditional energy knowledge are physics informed ML models and neuro-fuzzy systems further improves accuracy and reliability.

The impact of these techniques extends across the multiple industries and locations. Smart buildings benefit from optimized HVAC systems, automated lighting, and energy conservation strategies. In industrial settings, predictive analytics help businesses to anticipate the power demands , enhance and improves the efficiency even further to a great extent. Renewable energy systems that includes solar and wind power, leverage data-driven insights to maximize the output and maintain grid stability. Electric vehicles also rely on these models to optimize battery usage and charging cycles, ultimately improving energy efficiency.

Despite their promises, data-driven energy efficiency models often face challenges and Issues like data availability, real-time adaptability, and seamless integration with IoT and edge computing technologies remain hurdles to overcome. Addressing these challenges requires collaboration between researchers, industry professionals, and policymakers to create solutions that are scalable, interpretable, and cost-effective.

This paper explores the evolution of data-driven energy efficiency techniques, their applications, and upcoming future of sustainable energy management. With continuous advancements in AI and ML, these technologies hold the key to a more energy-efficient and environmentally conscious world.

II. Methodology

A. Data Parameters and Feature Analysis

The research examines a comprehensive set of building parameters that influence energy efficiency. Each parameter has been carefully selected based on its theoretical and practical significance in building energy performance:

1) Relative Compactness (X1): This parameter represents the ratio between the building's volume and external surface area normalized to a scale of 0 to 1. A higher value indicates a more compact building design, which

results in better energy efficiency due to reduced heat transfer surfaces.

2) Surface Area (X2): The surface area is measured in square meters. It includes all the surfaces in the building that are exposed to the external environment including walls, roofs, and even the ground floor. The surface area directly influences heat exchange between the building and its environment, making it a crucial factor in energy efficiency calculations.
3) Wall Area (X3): This parameter measures the total area of vertical surfaces in square meters, excluding windows. Wall area affects both thermal mass and heat transfer characteristics. The relationship between the wall area and other parameters, particularly the glazing area, showcases the building's thermal behavior.
4) Roof Area (X4): Measured in square meters, the roof area represents the uppermost building surface exposed to the sun rays and atmospheric conditions. This parameter is particularly significant for solar heat gain calculations and overall building energy performance, especially in regions with high exposure to the sun.
5) Overall Height (X5): Building height, measured in meters, significantly impacts natural ventilation patterns, the stack effect, and vertical temperature gradients. This parameter influences both heating and cooling loads by affecting air circulation and thermal stratification within the building.
6) Orientation (X6): The orientation is represented as a variable between the range of 1 to 4. It represents the building's primary front relative to cardinal points. This parameter significantly affects solar heat gain and natural lighting potential while affecting the heating and cooling energy requirements too.
7) Glazing Area (X7): Quantified as a percentage of the total window area, this parameter measures the proportion of transparent surfaces. The glazing area directly influences daylighting, solar heat gain, and thermal losses, making it an important factor in building energy performance.
8) Glazing Distribution (X8): Represented as a categorical variable (0-5), this parameter describes how window areas are distributed across different faces of the building. The window distribution around the building affects solar heat gain patterns and natural lighting distribution throughout the building.
9) Heating Load (Y1): The Heating load is measured in kilowatt-hours (kWh) and represents the annual energy requirement for maintaining comfortable indoor temperatures during cold weather conditions. The heating load calculation considers both steady-state heat transfer and dynamic thermal behaviors, making it a complex function of building geometry, material properties, and environmental conditions.
10) Cooling Load (Y2): The Cooling load is calculated in kilowatt-hours (kWh) and represents the annual energy requirement for maintaining comfortable indoor temperatures during warm weather conditions. The cooling load is mainly sensitive to solar orientation, glazing properties, and internal heat generation patterns, making it a critical indicator of building energy performance in cold climates.

The Heating and Cooling loads correlate with each other inversely, that is, an increase in any one of them relates to a decrease in the other. Building design and energy efficiency optimization must carefully take this trade-off into account. Dynamic thermal simulation techniques that take time-varying circumstances and thermal storage effects into consideration are used to calculate both loads.

B. Data Processing and Preparation

The data processing pipeline consists of several carefully designed steps to ensure data quality and analytical reliability:

1) Data Cleaning and Validation: The data processing starts by cleaning and validating the acquired data.

- Missing Value Treatment: The methodology employs a systematic approach to handling missing values:
  - Numerical features: Missing values are identified and evaluated for patterns
  - Statistical imputation using mean values for features with less than 5% missing data
  - Features with more than 5% missing data trigger a detailed investigation protocol
- Outlier Detection and Handling: The research implements a multi-step outlier detection process:
  - Z-score analysis with a threshold of ±3 standard deviations.
  - Interquartile Range (IQR) method with a factor of 1.5.
  - Modified Thompson Tau test for extreme value validation.
  - Domain-specific validation rules for each parameter
- Data Type Verification: Each parameter undergoes strict type checking and conversion:
  - Numerical parameters are validated for proper scaling and units.
  - Categorical variables are encoded using appropriate schemes.
  - Format consistency is enforced across all features.

2) Feature Engineering and Transformation: The methodology incorporates sophisticated feature engineering techniques:

1) Feature Scaling:
   - Standard scaling (z-score normalization) for numerical features.
   - Min-max scaling for bounded parameters.
   - Log transformation for heavily skewed distributions.
2) Feature Interaction Analysis:
   - Polynomial feature generation up to degree 2.
   - Interaction term creation for physically relevant parameter pairs.

- Principal Component Analysis (PCA) for dimensionality assessment.

C. Statistical Analysis Methods

1) Descriptive Statistics Implementation: The descriptive statistics framework employs specific computational methods:

- Central Tendency Measures:
  - Arithmetic mean with 95% confidence intervals
  - Weighted median for skewed distributions
  - Mode calculation with kernel density estimation
- Dispersion Metrics:
  - Standard deviation with bias correction
  - Variance computation with Bessel's correction
  - Coefficient of variation with percentage representation
- Distribution Shape Analysis:
  - Skewness calculation using the Fisher-Pearson coefficient
  - Kurtosis computation using Fisher's definition

2) Correlation Analysis Techniques: The correlation analysis employs multiple methods to ensure robust relationship detection:

- Pearson Correlation:
  - Computation of linear correlation coefficients
  - Significance testing with p-value thresholds of 0.05
  - Confidence interval estimation using Fisher's z-transformation
- Spearman Rank Correlation:
  - Non-parametric correlation assessment
  - Monotonic relationship detection
  - Rank-based coefficient calculation
- Mutual Information Analysis:
  - Information-theoretic relationship quantification
  - Non-linear dependency detection
  - Feature importance ranking based on mutual information scores

D. Feature Importance Analysis

The research implements a comprehensive feature importance analysis framework:

- Statistical Feature Importance:
  - ANOVA F-value calculation for numerical features
  - Chi-square test for categorical variables
  - Information gain ratio computation
- Model-Based Feature Importance:
  - Random Forest feature importance using mean decrease in impurity
  - XGBoost feature importance based on gain and coverage
  - Permutation importance calculation for model validation
- Feature Selection Methods:
  - Recursive feature elimination with cross-validation
  - Lasso regularization for feature selection
  - Stability selection with bootstrapped samples

E. Model Development and Validation

The modeling framework incorporates multiple algorithms with specific hyperparameter configurations:

- Linear Models:
  - Linear Regression with standardized coefficients
  - Ridge Regression ($\alpha = 1.0$, solver = $'\texttt{auto}'$)
  - Lasso Regression ($\alpha = 0.01, \max\_iter = 1000$)
- Ensemble Methods:
  - Random Forest (n_estimators=100, max_depth=None)
  - XGBoost (learning_rate=0.1, max_depth=6)
  - Gradient Boosting (n_estimators=100, learning_rate=0.1)
- Neural Networks:
  - Multi-layer Perceptron (hidden_layers=[100, 50], activation='relu')
  - Learning rate optimization using Adam optimizer
  - Dropout regularization (rate=0.2)

F. Validation and Performance Assessment

The methodology employs rigorous validation procedures:

- Cross-Validation Strategy:
  - K-fold cross-validation (k=5)
  - Stratified sampling for balanced fold creation
  - Repeated cross-validation for stability assessment
- Performance Metrics:
  - Mean Squared Error (MSE) for absolute error quantification
  - R-squared (R^2) for explained variance assessment
  - Mean Absolute Percentage Error (MAPE) for relative error measurement
- Model Comparison Framework:
  - Statistical significance testing of performance differences
  - Model ranking using multiple criteria
  - Ensemble model creation based on validation results

III. Model Development And Evaluation

A. Dataset

The Energy Efficiency dataset from the UCI Machine Learning Repository which has been used for our research purposes, is designed for analyzing energy consumption in buildings, comprising 768 samples that represent unique building configurations simulated using Ecotect software. The buildings take into consideration key parameters such as glazing area, glazing area distribution, orientation, relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution.

The primary aim of the dataset is to predict two continuous response variables: heating load, which refers to the energy required to heat the building, and cooling load, which denotes the energy required to cool it. These responses can also be rounded for multi-class classification purposes. This dataset facilitates the development of predictive models for heating and cooling loads, allows for the analysis of influential features on energy consumption, and aids in formulating strategies to minimize energy loads in residential buildings.

B. Metrics used

Analyzing a machine learning model guarantees that it generalizes well to unseen data and provides accurate predictions. In this study, we used 3 key evaluation metrics

- R-square-Goodness of Fit:It is also known as coefficient of determination which measures proportion of variance by the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (y_i - \widehat{y}_i)^2}{\sum_{i=1}^{m} (y_i - \overline{y})^2}$$

- Root Mean Squared Error(RMSE): Measure the average difference between the model's predicted values and actual values.

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \widehat{y}_i)^2}$$

- Cross-validation: A k-fold approach is implemented to ensure that there is no overfitting and the model is generalized across different data subsets.

Following are the obtained results for all the different evaluated models:

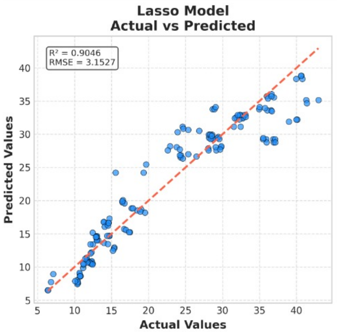


Fig. 1: Caption

C. Model Comparison

We used different machine learning models including Ridge, Lasso, Linear Regression, Random Forest, Neural Networks, and XGBoost to predict energy-efficient parameters accurately. Among the six models evaluated, XGBoost demonstrates the best overall performance with the highest R^2 value of 0.99852 and lowest RMSE of 0.39256, followed by Random Forest (R^2 = 0.99765) and neural network (R^2 = 0.99623). The traditional linear models (linear, ridge, and lasso regression) show relatively lower performance metrics, with R^2 values around 0.90-0.91 and significantly higher RMSE values above 3.0. This significant difference in performance suggests that the relationship between the energy efficiency parameters is highly nonlinear, which explains why tree-based ensemble methods (XGBoost and random forest) and neural networks can capture the underlying patterns more effectively. The consistently low CV standard deviation across all models (ranging from 0.000162 to 0.004489) indicates stable performance across different cross-validation folds. This indicates that regardless of the model choice, the predictions are bound to be reliable and consistent.

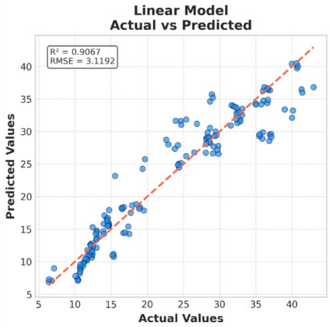


Fig. 2: Caption

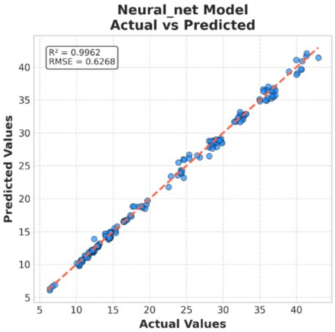


Fig. 3: Caption

D. Feature Importance Statistics

Model	R^2	RMSE	MAE	RAE	CV Mean	CV Std
linear	0.906655	3.119225	2.226601	0.240531	0.913292	0.004489
ridge	0.912130	3.026376	2.182907	0.235810	0.914495	0.004094
lasso	0.904642	3.152675	2.294545	0.247870	0.910354	0.003715
random_forest	0.997645	0.495452	0.351680	0.037991	0.997331	0.000589
xgboost	0.998522	0.392555	0.264183	0.028539	0.998359	0.000162
neural_net	0.996231	0.626767	0.476213	0.051443	0.994927	0.001917

TABLE I: Model Performance Comparison

TABLE II: Comprehensive Descriptive Statistics of Building Parameters

Parameter	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**	**Skew**	**Kurt**	**Mode**	**Range**	**IQR**	**CV**
Geometrical Parameters													
Relative Compactness	0.764	0.106	0.620	0.682	0.750	0.830	0.980	0.496	-0.707	0.620	0.360	0.148	13.842
Surface Area	671.708	88.086	514.500	606.375	673.750	741.125	808.500	-0.125	-1.059	514.500	294.000	134.750	13.114
Wall Area	318.500	43.626	245.000	294.000	318.500	343.000	416.500	0.533	0.117	294.000	171.500	49.000	13.697
Roof Area	176.604	45.166	110.250	140.875	183.750	220.500	220.500	-0.163	-1.777	220.500	110.250	79.625	25.575
Overall Height	5.250	1.751	3.500	3.500	5.250	7.000	7.000	0.000	-2.005	3.500	3.500	3.500	33.355
Design Parameters													
Orientation	3.500	1.119	2.000	2.750	3.500	4.250	5.000	0.000	-1.361	2.000	3.000	1.500	31.965
Glazing Area	0.234	0.133	0.000	0.100	0.250	0.400	0.400	-0.060	-1.328	0.100	0.400	0.300	56.841
Glazing Distribution	2.813	1.551	0.000	1.750	3.000	4.000	5.000	-0.089	-1.149	1.000	5.000	2.250	55.145
Energy Load Parameters													
Heating Load	22.307	10.090	6.010	12.993	18.950	31.675	43.100	0.360	-1.246	15.160	37.090	18.683	45.233
Cooling Load	24.588	9.513	10.900	15.620	22.080	33.133	48.030	0.396	-1.147	14.270	37.130	17.513	38.691

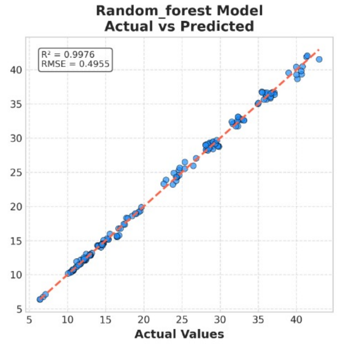


Fig. 4: Caption

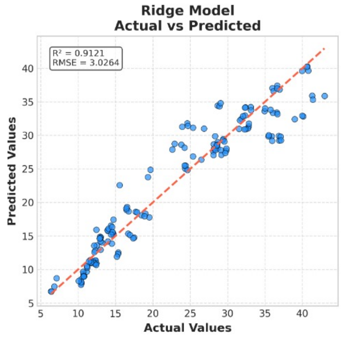


Fig. 5: Caption

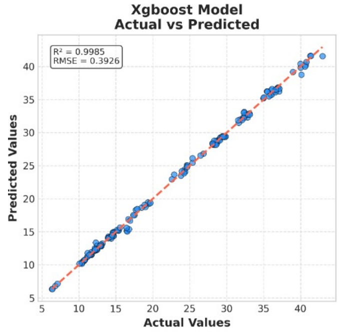


Fig. 6: Caption

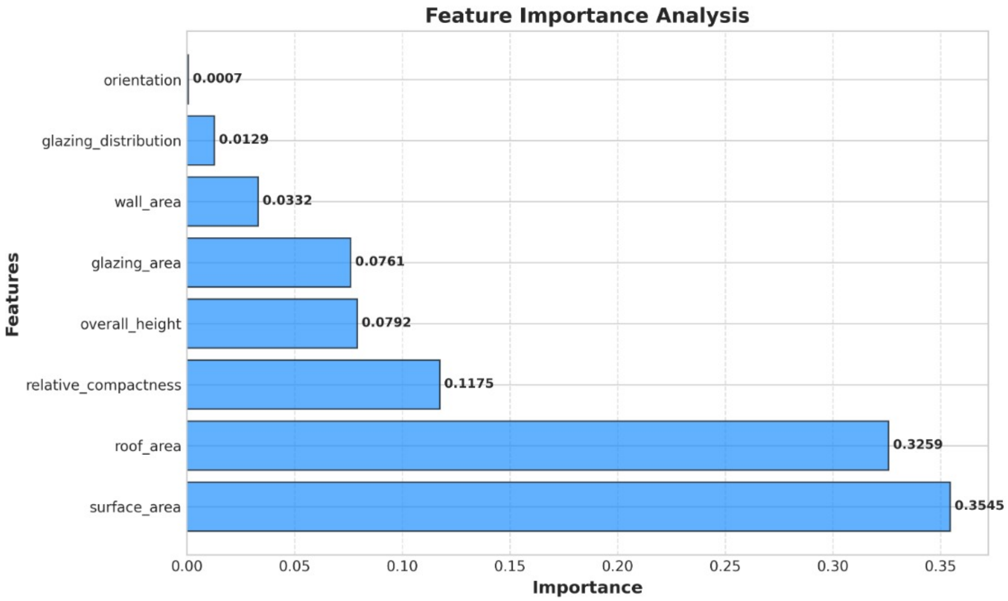
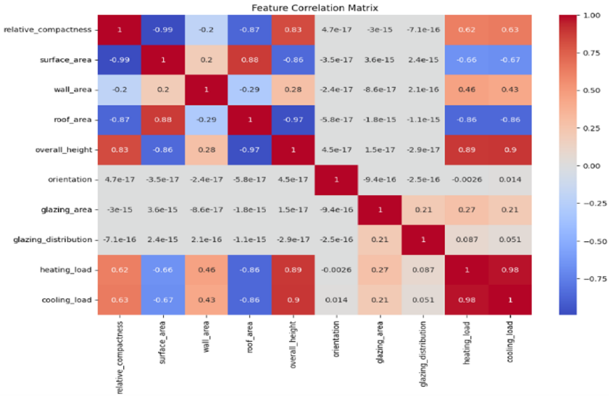


Fig. 7: Caption



	relative_compactness	surface_area	wall_area	roof_area	overall_height	orientation	glazing_area	glazing_distribution	heating_load	cooling_load
relative_compactness	1	-0.99	-0.2	-0.87	0.83	4.7e-17	-3e-15	-7.1e-16	0.62	0.63
surface_area	-0.99	1	0.2	0.88	-0.86	-3.5e-17	3.6e-15	2.4e-15	-0.66	-0.67
wall_area	-0.2	0.2	1	-0.29	0.28	-2.4e-17	-8.6e-17	2.1e-16	0.46	0.43
roof_area	-0.87	0.88	-0.29	1	-0.97	-5.8e-17	-1.8e-15	-1.1e-15	-0.86	-0.86
overall_height	0.83	-0.86	0.28	-0.97	1	4.5e-17	1.5e-17	-2.9e-17	0.89	0.9
orientation	4.7e-17	-3.5e-17	-2.4e-17	-5.8e-17	4.5e-17	1	-9.4e-16	-2.5e-16	-0.0026	0.014
glazing_area	-3e-15	3.6e-15	-8.6e-17	-1.8e-15	1.5e-17	-9.4e-16	1	0.21	0.27	0.21
glazing_distribution	-7.1e-16	2.4e-15	2.1e-16	-1.1e-15	-2.9e-17	-2.5e-16	0.21	1	0.087	0.051
heating_load	0.62	-0.66	0.46	-0.86	0.89	-0.0026	0.27	0.087	1	0.98
cooling_load	0.63	-0.67	0.43	-0.86	0.9	0.014	0.21	0.051	0.98	1

Fig. 8: Caption

E. Comparison with similar models in the past

When we compare our best-performing model (XGboost) with similar models used in the past for similar tasks utilizing different datasets, we can see similar results for the RMSE score and an improvement in the R^2 score. Our model considers several different parameters that will affect a building's energy utilization and not just a single parameter, making it more diverse and unique.

TABLE III: Comparison of Energy Efficiency Metrics

Paper Number	**RMSE**	**R^2**
Using a Surrogate Model to Analyze the Impact of Geometry on Energy Efficiency of Buildings	0.022	0.999
AI-based Forecasting for Optimised Solar Energy Management and Smart Grid Efficiency	0.85	0.95
An Efficient Machine Learning Based Optimization Framework to Analyse the Short-Term Solar Energy	0.12	0.98
AI-based Optimization Framework for Short-Term Solar Energy Prediction	0.15	0.97
Proposed Research	**0.3926**	**0.9985**

References

[1] A. J. Gutiérrez Trashorras, J. M. González-Caballín Sánchez, E. Álvarez Álvarez and J. P. Paredes Sánchez, "Certification of Energy Efficiency in New Buildings: A Comparison Among the Different Climatic Zones of Spain," in IEEE Transactions on Industry Applications, vol. 51, no. 4, pp. 2726-2731, July-Aug. 2015, doi: 10.1109/TIA.2015.2394374.

[2] A. M. Al-Ghaili, H. Kasim, N. M. Al-Hada, M. Othman and M. A. Saleh, "A Review: Buildings Energy Savings - Lighting Systems Performance," in IEEE Access, vol. 8, pp. 76108-76119, 2020, doi: 10.1109/ACCESS.2020.2989237.

[3] E. Mills, "Action-Oriented Energy Benchmarking for Nonresidential Buildings," in Proceedings of the IEEE, vol. 104, no. 4, pp. 697-712, April 2016, doi: 10.1109/JPROC.2016.2520638.

[4] J. Pan, R. Jain, S. Paul, T. Vu, A. Saifullah and M. Sha, "An Internet of Things Framework for Smart Energy in Buildings: Designs, Prototype, and Experiments," in IEEE Internet of Things Journal, vol. 2, no. 6, pp. 527-537, Dec. 2015, doi: 10.1109/JIOT.2015.2413397..

[5] A. Pellegrino, V. R. M. Lo Verso, L. Blaso, A. Acquaviva, E. Patti and A. Osello, "Lighting Control and Monitoring for Energy Efficiency: A Case Study Focused on the Interoperability of Building Management Systems," in IEEE Transactions on Industry Applications, vol. 52, no. 3, pp. 2627-2637, May-June 2016, doi: 10.1109/TIA.2016.2526969.

[6] O. Jogunola, C. Morley, I. J. Akpan, Y. Tsado, B. Adebisi and L. Yao, "Energy Consumption in Commercial Buildings in a Post-COVID-19 World," in IEEE Engineering Management Review, vol. 50, no. 1, pp. 54-64, 1 Firstquarter,march 2022, doi: 10.1109/EMR.2022.3146591.

[7] R. Loggia, A. Flamini, A. Massaccesi, C. Moscatiello and L. Martirano, "A Case Study of a Renovation of a Historical University Department: The Nearly Zero-Energy Refurbished Buildings," in IEEE Transactions on Industry Applications, vol. 58, no. 6, pp. 6970-6980, Nov.-Dec. 2022, doi: 10.1109/TIA.2022.3195823.

[8] L. Bereketeab, A. Zekeria, M. Aloqaily, M. Guizani and M. Debbah, "Energy Optimization in Sustainable Smart Environments With Machine Learning and Advanced Communications," in IEEE Sensors Journal, vol. 24, no. 5, pp. 5704-5712, 1 March1, 2024, doi: 10.1109/JSEN.2024.3355229.

[9] L. Angel Iturralde Carrera et al., "Integration of Energy Management and Efficiency System for Buildings With Zero Carbon Emissions: A Case of Study," in IEEE Access, vol. 12, pp. 64237-64251, 2024, doi: 10.1109/ACCESS.2024.3396816.

[10] A. N. Khan et al., "Dynamic Temporal Analysis and Modeling of Residential Lighting Consumption for Energy Efficiency and Sustainability," in IEEE Access, vol. 12, pp. 154365-154380, 2024, doi: 10.1109/ACCESS.2024.3467337.

[11] Bouquet, P., Jackson, I., Nick, M., Kaboli, A. (2023). AI-based forecasting for optimised solar energy management and smart grid efficiency. International Journal of Production Research, 62(13), 4623–4644.

[12] Bhumika Bhatta , Ralph Evins , Paul Westermann. (2021). Using a surrogate model to analyze the impact of geometry on energy efficiency of buildings. Proceedings of Building Simulation: 17th Conference of IBPSA.

[13] Parida, Raj Roy, Monideepa Parida, Ajaya Khan, AsifUddin Sahoo, Biswa. (2024). An Efficient Machine Learning Based Optimization Framework to Analyse the Short-Term Solar Energy. 1-6. 10.1109/ICCSC62048.2024.10830308.

[14] ash, P K Mishra, Shaktinarayana Tripathy, L. Satapathy, Prachitara Sahani, Nitasha. (2020). An Efficient Machine Learning Approach for Accurate Short Term Solar Power Prediction. 10.48550/arXiv.2003.12088.