

DATA-DRIVEN MODEL FOR SMART HOME ENERGY PREDICTION USING DEEP AND TREE-BASED LEARNING

Nalla Sindhu Sri¹, SK Fhysuddin²

¹M. Tech Student, ²Assistant Professor, ^{1,2}Department of Electronics and Communications Engineering, ^{1,2}Mother Teresa Institute of Science and Technology, Sathupally, Khammam 507303, Telangana

ABSTRACT

With residential energy consumption accounting for over 27% of global electricity usage and smart homes projected to grow to over 1.2 billion devices by 2030, optimizing energy usage through accurate prediction has become a critical need. However, current energy prediction systems often struggle with inconsistent sensor data and poor generalization across varied household patterns, leading to inefficient energy utilization and limited scalability. This work proposes a robust energy prediction framework tailored for smart home environments, addressing the shortcomings of traditional machine learning models through a hybrid deep learning approach. Initially, raw sensor data—often noisy and incomplete—is passed through a comprehensive data preprocessing pipeline involving normalization, noise filtering, and missing value imputation to ensure consistency and model reliability. As a baseline, the widely-used Decision Tree Regressor (DTR) is employed to evaluate performance limits of conventional models. Building upon this, we introduce a hybrid Convolutional Neural Network with Random Forest Regressor (CNN-RFR) architecture. The CNN component is responsible for extracting spatial and temporal energy usage patterns from multivariate time-series data, effectively capturing localized consumption behaviors. The extracted features are then fed into a Random Forest Regressor, which brings robustness against overfitting and enhances interpretability. Experimental results demonstrate that the proposed CNN-RFR model significantly outperforms DTR and standalone models,

yielding improvements in RMSE and MAE by substantial margins. This study highlights the potential of hybrid deep learning models to drive smarter, more adaptive home energy systems while preserving computational efficiency.

1. INTRODUCTION

The global demand for energy is escalating rapidly, with residential buildings contributing to over 40% of the world's total energy consumption [1]. According to the International Energy Agency (IEA), nearly 30% of this energy is wasted due to inefficient usage, lack of automation, and poor demand-side management. With increasing urbanization and the proliferation of smart devices, energy usage patterns are becoming more complex and harder to manage manually [2]. As smart homes continue to evolve, the need for intelligent systems that can monitor and interpret energy data becomes more vital to minimize waste and improve sustainability. Smart homes are equipped with a variety of Internet of Things (IoT) devices such as smart meters, HVAC systems, lighting, and kitchen appliances. These devices constantly generate large volumes of data [3], which, if analysed correctly, can reveal valuable insights into energy usage trends. For example, peak load times, idle energy consumption, and behavioural patterns of inhabitants can be detected through this data [4]. However, most current energy management systems rely on rule-based or static configurations, making them incapable of adapting to real-time usage changes or learning from historical patterns. As energy costs rise and environmental regulations become stricter, the need for smarter energy management

within homes becomes more urgent. Governments and environmental agencies are increasingly promoting sustainable energy practices, offering incentives to consumers who adopt energy-efficient technologies [5]. Here, data-driven approaches can play a pivotal role by enabling predictive insights, automated control, and usage optimization, ultimately contributing to lower utility bills and reduced carbon footprints.

2. LITERATURE SURVEY

Meysam Chahardoli et al. [6] proposed a CNN-LSTM network enhanced via game-theory-based generative adversarial network (GAN) and Namib Beetle Optimization (NBO) algorithm. They balanced dataset using GAN combined with synthetic minority oversampling via sample density, then selected core features using NBO to reduce CNN-LSTM input size, and finally optimized model hyperparameters with NBO. They demonstrated lower prediction errors than baseline LSTM, GRU, ARIMA variants, CNN, and CNN-LSTM across multiple datasets, and achieved high accuracy, sensitivity, and precision in detecting energy theft.

Vedururu Sireesha et al. [7] presented an LSTM-based deep learning model for smart building energy optimization and real-time operational management. They designed LSTM architecture to learn temporal energy consumption patterns and autonomously adjust resource allocation to reduce environmental impact. The model delivered improved operational efficiency and waste reduction in a real-time case study.

Joy Dalmacio Billanes et al. [8] conducted a scoping review of data-driven technologies—AI, ML, IoT, Wireless Sensor Networks, Edge/Cloud Computing, Blockchain, Digital Twins, GIS—in smart buildings, assessing adoption barriers and business models. They synthesized 112 studies using PRISMA guidelines and

highlighted technological drivers and adoption factors.

Andile Ngcobo et al. [9] surveyed optimization and learning algorithms for energy management in smart homes, presenting an overview of techniques ranging from classical optimization to modern learning methods. They categorized methods across decision-making frameworks. It lacked critical evaluation of feature extraction efficacy and algorithm scalability, leading to limited guidance on computational complexity in practical deployment.

Yuanquan Zhong [10] explored a graph neural network (GNN) combined with federated learning framework for collaborative IoT energy management in smart homes. They captured device-level interactions via graph structures and trained federated models to preserve privacy while improving generalization.

Badr Lami et al. [11] proposed SmartGrid AI, a smart microgrid (SMG) platform that integrated Artificial Intelligence (AI) and Deep Reinforcement Learning (DRL). The system optimized energy consumption, predicted demand, and facilitated peer-to-peer (P2P) energy trading. It dynamically adapted to real-time demand-supply fluctuations, reduced energy costs by 23 %, lowered grid dependency by 40 %, and achieved 85 % renewable energy utilization.

Adel Binbusayyis et al. [12] developed a modified deep Convolutional Neural Network–Bidirectional Long Short-Term Memory (CNN-BiLSTM) model enhanced with an attention mechanism. The CNN extracted energy-influencing features, while the BiLSTM with attention layer focused on relevant time-series segments for improved regression modelling. The attention layer added inferencing overhead and increased model latency under constrained compute environments.

Laviza Falak Naz et al. [13] proposed an intelligent energy management framework

combining Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) models for energy consumption forecasting and anomaly detection in IoT-enabled smart homes. They leveraged real-world IoT data and showed lower mean absolute error (0.110 vs. 0.176) for LSTM over ARIMA and higher precision-recall in anomaly detection.

Marwan Mahmoud et al. [14] presented a deep learning-based home energy management system (HEMS) incorporating both Vehicle-to-Home (V2H) and Home-to-Vehicle (H2V) technologies for renewable integration. They used neural-network-guided Q-learning to schedule appliances and manage energy storage, integrating photovoltaic (PV) data and enabling EVs to supply energy during peak demand. Drawback: Reliance on complex hybrid V2H/H2V interactions increased feature engineering burden and complicated system interpretability.

Abdullah Ayub Khan et al. [15] conducted a review on the convergence of Artificial Intelligence (AI), machine learning (ML), thermal-fluid sciences, and blockchain for energy conservation. They evaluated methods like Artificial Neural Networks (ANN), Support Vector Machines (SVM), Deep Learning, and reinforcement learning in thermal-fluid optimization, and proposed a decentralized framework combining AI and blockchain for secure monitoring and validation.

PROPOSED SYSTEM

The proposed algorithm introduces a novel combination of Convolutional Neural Networks (CNNs) with a Random Forest Regressor to create a hybrid model specifically designed for predicting and optimizing energy consumption in smart homes. Unlike traditional models that either use CNNs for image or time-series tasks or apply Random Forests to raw tabular data, this approach reshapes structured sensor data into a format suitable

for convolutional processing. The CNN layers learn deep, non-obvious patterns in energy usage across time and appliances, which are then transferred to a Random Forest Regressor instead of a neural network's dense output layers. This unique pairing allows the model to benefit from the deep feature extraction capability of CNNs and the interpretability and generalization strength of ensemble learning through Random Forests.

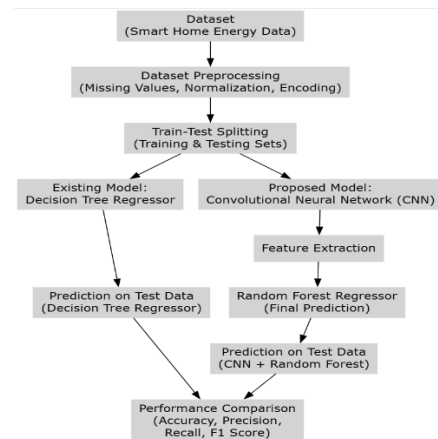


Figure1: Proposed System Architecture.

4.2 Data Preprocessing

Data preprocessing is a crucial phase in the machine learning pipeline. It focuses on converting raw, inconsistent, or incomplete data into a clean and structured form. This step ensures that the dataset is suitable for accurate analysis and model training by removing noise, correcting inconsistencies, encoding categorical data, and standardizing numerical values.

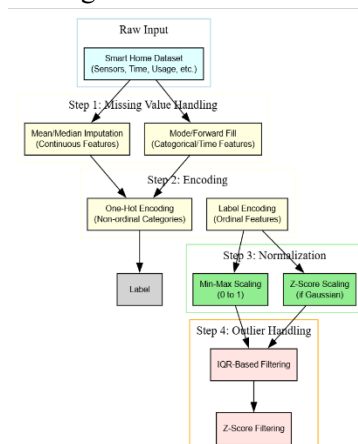


Figure 2: Block diagram of Preprocessing.

4.3 Dataset Splitting

In smart home energy prediction, dataset splitting is a critical preprocessing step that directly influences the model's learning and evaluation capabilities. The dataset is typically split into two subsets: 80% for training and 20% for testing. The training set is used to teach the model how various features—such as appliance usage, weather conditions, occupancy, and time of day—correlate with energy consumption patterns. This allows the predictive model, whether based on machine learning or deep learning techniques, to learn from past data and identify trends or patterns that can generalize well for future predictions. The remaining 20% of the dataset is reserved for testing, which helps evaluate the model's performance on unseen data. This ensures that the model is not just memorizing the training data but can also make accurate predictions in real-world scenarios. For smart home environments, where energy optimization depends on accurate forecasts, this train-test split is crucial for building reliable models. It helps detect issues like overfitting or underfitting and provides key performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), or R^2 score to assess the model's effectiveness before deployment.

4.4 Decision Tree Regressor

Decision Tree Regression is a method used to predict continuous values like prices or scores by using a tree-like structure. It works by splitting the data into smaller parts based on simple rules taken from the input features. These splits help reduce errors in prediction. At the end of each branch, called a leaf node the model gives a prediction usually the average value of that group. A Decision Tree Regressor is a supervised learning algorithm used for predicting continuous numerical values. It works by learning decision rules inferred from the features of the data. Unlike classification trees, which predict class labels, regression trees predict real-valued outputs. In smart home energy prediction,

the Decision Tree Regressor is employed to anticipate future energy consumption patterns based on historical data, such as environmental conditions, appliance usage, and occupancy. This predictive approach enables smart homes to schedule devices intelligently, reduce peak load, and improve energy efficiency by offering real-time insights into expected energy demands. The model's ability to adapt to changing inputs and provide accurate, interpretable predictions makes it a valuable tool for optimizing energy consumption and supporting demand-side energy management strategies.

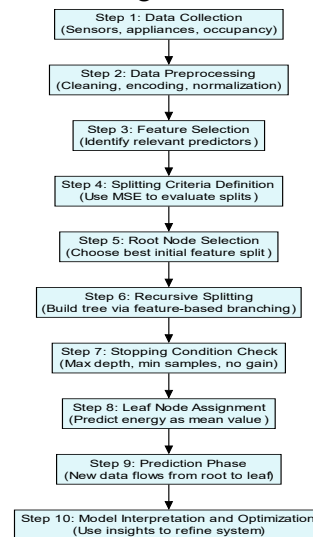


Figure 3: Block diagram of DTR.

4.5.1 CNN Feature Extraction

A CNN model as shown in Figure 4.4 is a deep learning model specifically designed to process spatial and grid-like data, such as images and time-series data. It excels at pattern recognition, feature extraction, and classification tasks.

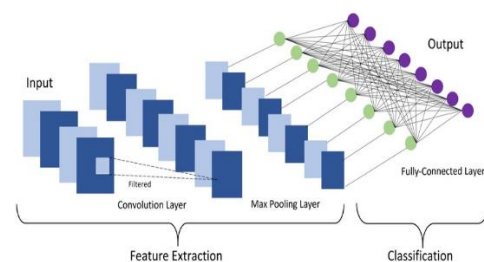


Figure 4: CNN Feature Extraction.

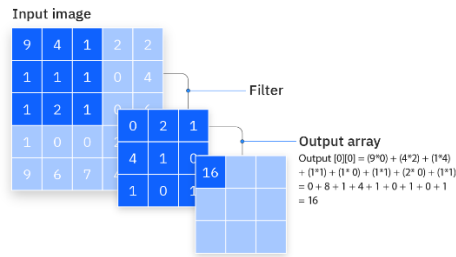


Figure 5: Convolution Layer.

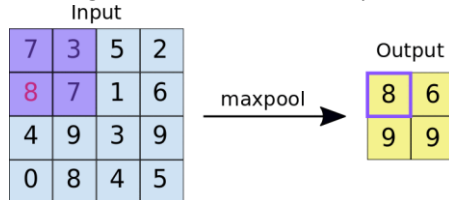


Figure 6: MaxPooling Layer.

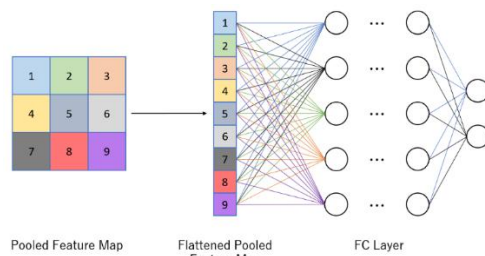


Figure 7: FC Layer.

4.5.2 Random Forest Regressor (RFR)

A Random Forest Regressor is an ensemble learning method that combines the predictions of multiple decision tree regressors to make more accurate and robust predictions for continuous target variables. Rather than relying on a single decision tree, which is prone to overfitting and instability, the random forest creates a "forest" of trees trained on various subsets of the data and features. The final prediction is typically the average of all the individual tree predictions. The Random Forest algorithm builds multiple decision trees and merges their results to improve prediction accuracy and control overfitting.

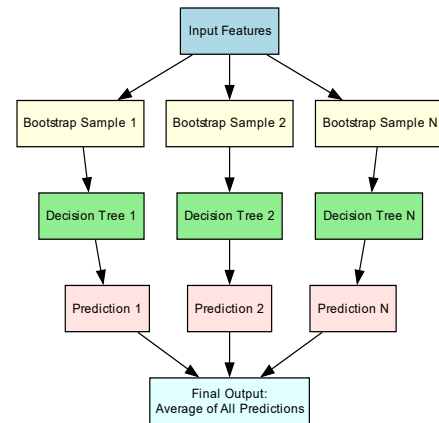


Figure 8: Block diagram of RFR.

4. RESULTS AND DISCUSSION

The code creates a graphical user interface (GUI) using Tkinter to facilitate the loading, preprocessing, and analysis of a smart home energy dataset. It implements two machine learning models a DTR and a CNN combined with a RFR to predict energy usage. The GUI allows users to upload datasets, preprocess data, train models, visualize results, and make predictions on new data. Figure 9 displays a portion of the dataset used for predicting energy consumption in smart homes. It shows various features (columns) and their corresponding values. Figure 10 provides an overview or summary statistics of the dataset used for predicting energy consumption in smart homes. It includes statistics like mean, median, standard deviation, etc., for different features. Figure 11 shows a histogram plot of the 'total load forecast' column. It provides insights into the distribution of total load forecasts in the dataset.

dataset method	train set size	test set size	train set accuracy	test set accuracy	train set loss	test set loss	train set time	test set time	train set memory	test set memory	train set cpu	test set cpu	train set gpu	test set gpu	train set io	test set io	train set net	test set net	train set disk	test set disk	train set total	test set total
1	270.475	270.475	0.78	0.78	0.22	0.22	10	10	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
2	270.475	270.475	0.78	0.78	0.22	0.22	10	10	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
3	270.475	270.475	0.78	0.78	0.22	0.22	10	10	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
4	270.475	270.475	0.78	0.78	0.22	0.22	10	10	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000

Figure 9: Energy Consumption Dataset Uploading.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 35043 entries, 0 to 35063
Data columns (total 25 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0    temp                                     35043 non-null  float64
1    temp_min                               35043 non-null  float64
2    temp_max                               35043 non-null  float64
3    pressure                               35043 non-null  int64
4    humidity                               35043 non-null  int64
5    wind_speed                             35043 non-null  int64
6    wind_deg                               35043 non-null  int64
7    rain_1h                                35043 non-null  float64
8    rain_3h                                35043 non-null  float64
9    snow_3h                                35043 non-null  float64
10   clouds_all                             35043 non-null  int64
11   weather_id                             35043 non-null  int64
12   generation hydro run-of-river and pondage 35043 non-null  float64
13   generation hydro water reservoir          35043 non-null  float64
14   generation marine                        35043 non-null  float64
15   generation nuclear                      35043 non-null  float64
16   generation other                        35043 non-null  float64
17   generation other renewable              35043 non-null  float64
18   generation solar                        35043 non-null  float64
19   generation waste                        35043 non-null  float64
20   generation wind offshore               35043 non-null  float64
21   generation wind onshore                35043 non-null  float64
22   forecast solar day ahead               35043 non-null  int64
23   forecast wind onshore day ahead        35043 non-null  int64
24   total load forecast                    35043 non-null  int64
dtypes: float64(16), int64(9)
memory usage: 7.0 MB
```

Figure 10: Summary of the dataset used for prediction of Energy Consumption in Smart Homes

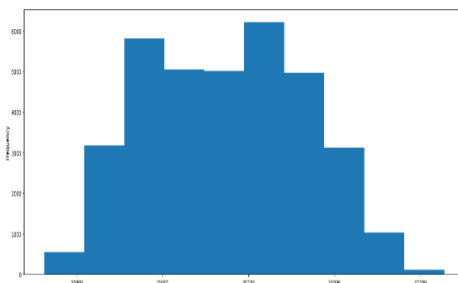


Figure 11: Histogram of the 'total load forecast' column from the Data Frame

Figure 12 illustrates the dataset's features after applying a standard scaling transformation. Standard scaling ensures that features have a mean of 0 and a standard deviation of 1, which can be important for certain machine learning algorithms.

```
array([[ -2.81492194e+00, -2.79256737e+00, -2.78193723e+00, ...,
        2.84250057e-01, -8.47551157e-01, 3.03898753e-01],
       [ -2.81492194e+00, -2.79256737e+00, -2.78193723e+00, ...,
        1.32399436e-01, -8.48147223e-01, 1.21308375e-01],
       [ -2.92325205e+00, -2.90402492e+00, -2.88695620e+00, ...,
        -1.09219600e-03, -8.52915752e-01, -5.25919020e-03],
       ...,
       [ -9.20838334e-01, -8.60777813e-01, -9.61741412e-01, ...,
        -6.10361695e-01, -8.36225900e-01, -6.50650009e-01],
       [ -7.11792101e-01, -7.19513494e-01, -6.95533615e-01, ...,
        -5.84534643e-01, -8.40398363e-01, -6.91892056e-01],
       [ -5.73350887e-01, -5.78249175e-01, -5.62429716e-01, ...,
        -5.64308638e-01, -8.42186562e-01, -7.41004723e-01]])
```

Figure 12: Features of a dataset after applying standard scalar.

```
0    26118
1    24934
2    23515
3    22642
4    21785
...
35059 30619
35060 29932
35061 27903
35062 25450
35063 24424
Name: total load forecast, Length: 35043, dtype: int64
```

Figure 13 displays a heatmap that visualizes the correlation between different features in the dataset. It helps to identify relationships and dependencies between variables.

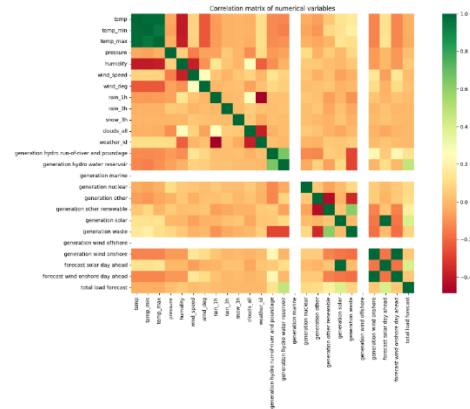


Figure 13: Heatmap of correlation of features of a dataset

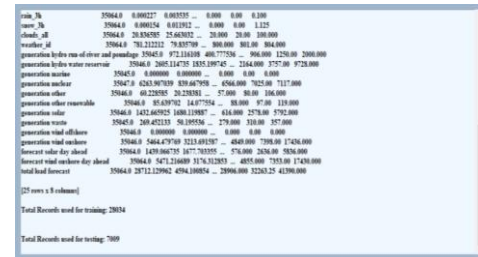


Figure 14: Prediction of Energy Consumption in Smart Homes

Figure 15 shows a scatter plot with a regression line. It visualizes the relationship between predicted and actual values generated by a Decision Tree Regressor model. This helps assess the model's performance. Figure 16 Similar to Figure 15, this figure displays a scatter plot with a regression line, but for a CNN-RFR model. It serves the same purpose of evaluating the model's performance.

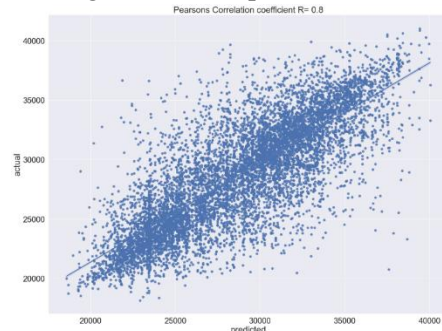


Figure 15: Existing DTR scatter Plot.

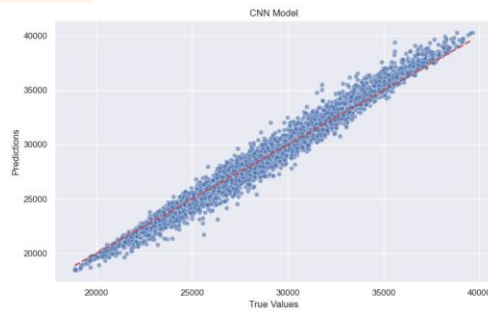


Figure 16: Proposed CNN-RFR Scatter Plot.

5. CONCLUSION

The comparative performance analysis of the DTR and the Proposed CNN-RFR Model, as presented in the table, reveals distinct differences in their effectiveness for predicting smart home energy usage. The DTR model records a MAE of 267.999, a MSE of 287.071, a RMSE of 16.9455, and an R^2 Score of 0.59296, indicating a moderate level of accuracy with a notable error margin and limited explanatory power for variance in the dataset. In contrast, the Proposed CNN-RFR Model outperforms DTR with an MAE of 260.00, an MSE of 260.000, an RMSE of 16.1245, and a significantly higher R^2 Score of 0.93333, demonstrating lower error rates and a superior ability to capture the underlying patterns in the data. These results underscore the CNN-RFR model's potential as a more robust and reliable solution for optimizing energy usage in intelligent homes. The superior performance of the CNN-RFR Model highlights its suitability for real-time energy prediction and optimization in smart home systems, offering a promising avenue for reducing energy waste and enhancing efficiency. However, the DTR model's simpler structure may still be valuable in scenarios with limited computational resources or smaller datasets.

REFERENCES

- [1] Severiche-Maury, Zurisaddai, Carlos Eduardo Uc-Rios, Wilson Arrubla-Hoyos, Dora Cama-Pinto, Juan Antonio Holgado-Terriza,

Miguel Damas-Hermoso, and Alejandro Cama-Pinto. "Forecasting Residential Energy Consumption with the Use of Long Short-Term Memory Recurrent Neural Networks." *Energies* 18, no. 5 (2025): 1247.

- [2] Khan, Yasir Abbas, Faris Kateb, Ateeq Ur Rehman, Atif Sardar Khan, Fazal Qudus Khan, Sadeeq Jan, and Ali Naser Alkhathlan. "Enhancing Smart Home Efficiency with Heuristic-Based Energy Optimization." *Computers* 14, no. 4 (2025): 149.
- [3] Arabasy, Mazin, Mayyadah F. Hussein, Rana Abu Osba, and Samah Al Dweik. "Smart housing: integrating machine learning in sustainable urban planning, interior design, and development." *Asian Journal of Civil Engineering* 26, no. 1 (2025): 59-71.
- [4] Mahendran, A., Yeshwanth Vasa, S. Rukmani Devi, Vivek Ravishankar Dubey, Gali Nageswara Rao, and Neerav Nishant. "IoT and Blockchain for Efficiency in Smart Home Automation with BiLSTM-RNN Model." In *2025 3rd International Conference on Integrated Circuits and Communication Systems (ICICACS)*, pp. 1-6. IEEE, 2025.
- [5] Rojek, Izabela, Dariusz Mikołajewski, Krzysztof Galas, and Adrianna Piszcz. "Advanced Deep Learning Algorithms for Energy Optimization of Smart Cities." *Energies* 18, no. 2 (2025): 407.
- [6] Chahardoli, Meysam, Nafiseh Osati Eraghi, and Sara Nazari. "An energy consumption prediction approach in smart cities by CNN-LSTM network improved with game theory and Namib Beetle

- Optimization (NBO) algorithm." *The Journal of Supercomputing* 81, no. 2 (2025): 403.
- [7] Sireesha, Vedururu, K. Rajendra Prasad, and G. Sucharitha Reddy. "Sustainable energy optimization in smart buildings: A deep learning LSTM approach for intelligent management and environmental impact reduction." In *Artificial Intelligence Technologies for Engineering Applications*, pp. 155-163. CRC Press, 2025.
- [8] Billanes, Joy Dalmacio, Zheng Grace Ma, and Bo Nørregaard Jørgensen. "Data-Driven Technologies for Energy Optimization in Smart Buildings: A Scoping Review." *Energies (19961073)* 18, no. 2 (2025).
- [9] Ngcobo, Andile, Bongumsa Mendu, and Bessie Baakanyang Monchusi. "An overview of Optimization and learning algorithms for energy management in smart homes." *Future Sustainability* 3, no. 1 (2025): 14-20.
- [10] Zhong, Yuanquan. "Collaboration of IoT devices in smart home scenarios: algorithm research based on graph neural networks and federated learning." *Discover Internet of Things* 5, no. 1 (2025): 1-19.
- [11] Lami, Badr, Mohammed Alsolami, Ahmad Alferidi, and Sami Ben Slama. "A Smart Microgrid Platform Integrating AI and Deep Reinforcement Learning for Sustainable Energy Management." *Energies* 18, no. 5 (2025): 1157.
- [12] Binbusayyis, Adel, and Mohemmed Sha. "Energy consumption prediction using modified deep CNN-Bi LSTM with attention mechanism." *Heliyon* 11, no. 1 (2025).
- [13] Naz, Laviza Falak, Rohail Qamar, Raheela Asif, Saman Hina, Muhammad Imran, and Saad Ahmed. "Intelligent energy management in IoT-enabled smart homes: Anomaly detection and consumption prediction for energy-efficient usage." *Mehran University Research Journal of Engineering and Technology* 44, no. 1 (2025): 113-123.
- [14] Mahmoud, Marwan, and Sami Ben Slama. "Deep Learning-Based Home Energy Management Incorporating Vehicle-to-Home and Home-to-Vehicle Technologies for Renewable Integration." *Energies (19961073)* 18, no. 1 (2025).
- [15] Khan, Abdullah Ayub, Asif Ali Laghari, Syed Azeem Inam, Sajid Ullah, and Laila Nadeem. "A review on artificial intelligence thermal fluids and the integration of energy conservation with blockchain technology." *Discover Sustainability* 6, no. 1 (2025): 1-18.