Machine Learning–Based Carbon Emission and Energy-Use Prediction for Sustainable Living

Tapodyuti Sarkar
Student of Master's in Technology

Department of Computational Intelligence
SRM Institute of Science and Technology
Kattankulathur, Tamil Nadu - 603203, India
Email: ts2972@srmist.edu.in

Dr. B. Pitchaimanickam
Department of Computational Intelligence
SRM Institute of Science and Technology
Kattankulathur, Tamil Nadu - 603203, India
Email: bpitmani@srmist.edu.in

***Abstract*— Climate change driven by rising carbon dioxide emissions is a critical global threat, directly addressed by SDG-13: Climate Action. Residential and small-scale energy use accounts for a significant share of global emissions, yet individuals lack simple, data-driven tools to understand and reduce their footprint. This project proposes a machine learning framework to predict household energy consumption and associated CO_2 emissions from easily measurable features such as appliance usage, fuel type, and basic building characteristics. The system compares multiple models (Linear Regression, Random Forest, Gradient Boosting, XGBoost, and a lightweight neural network) using metrics like MAE, RMSE, and R^2, and selects the best model to power a user-facing web dashboard. The expected outcome is an interpretable prediction tool that not only estimates emissions but also provides feature-based insights and simple what-if scenarios (e.g., switching fuels, reducing usage) to support practical decarbonization decisions at the household level. This contributes to SDG-13 by translating ML-based carbon-emission research into an accessible decision-support application.**

Keywords*— *Machine Learning, Carbon Emission Prediction, Energy Consumption, Sustainable Living, SDG-13, Random Forest, XGBoost, Interpretability

I. INTRODUCTION

Climate change is driven largely by human greenhouse gas emissions, with carbon dioxide from energy use as the dominant contributor. Residential and commercial buildings account for a substantial share of global final energy consumption and related CO_2 emissions, directly linking this sector to SDG-13: Climate Action. In many regions, especially developing countries, increasing adoption of energy-intensive appliances and rising electricity/LPG use intensify the climate impact of households.

Most individuals lack quantitative tools that convert their everyday energy-use patterns into estimated emissions and clear, personalized reduction strategies, creating a gap between awareness and action. Recent research shows that machine learning models can accurately predict building energy demand and CO_2 emissions from features such as floor area, building type, climate zone, and usage profiles. However, many existing ML solutions focus on aggregate forecasts (city or national level) or act as complex black boxes, limiting their usefulness for household-level decision making and practical climate-action planning.

This work addresses these gaps by:

1. Developing a comparative ML framework that evaluates multiple regression and ensemble models for household-level CO_2 and energy prediction.
2. Providing interpretable outputs through feature-importance analysis and SHAP-based explanations to identify emission drivers.
3. Deploying the best-performing model in a simple web dashboard where users can input their data and receive actionable recommendations aligned with SDG-13 goals.

The paper is organized as follows: Section II reviews related literature on ML-based carbon and energy prediction. Section III describes the problem formulation and proposed methodology. Section IV outlines the dataset and experimental setup. Section V presents preliminary results and discussion. Section VI concludes with future directions.

II. RELATED WORK

Recent advances in machine learning have significantly improved the accuracy of carbon emission and energy-consumption prediction at various scales. This section surveys key studies that inform the design of this project.

A. National and Aggregate-Scale Prediction

Ye et al. (2025) evaluated 14 models including ARMA, Random Forest, XGBoost, and CNN–RNN hybrids for daily national CO_2 emissions, achieving R^2 values between 0.714 and 0.932. They recommended ensemble ML as the best trade-off between accuracy and complexity. However, their work focused on national-level data and did not address household or building-level tools. Similarly, Ahmed et al. (2023) used regression, SVR, Random Forest, and Gradient Boosted Trees for forecasting U.S. national CO_2 emissions, but did not connect predictions to household actions or behavior change.

B. Residential and Building-Level Prediction

Khan et al. (2025) compared Decision Tree, Random Forest, Ridge Regression, Gradient Boosting, SVR, XGBoost, and LSTM for predicting daily residential CO_2 emissions in U.S. buildings (2022–2024). Tuned ensembles and LSTM achieved the best RMSE and R^2, outperforming statistical baselines. However, their work used aggregated national datasets and lacked interactive tools or per-household explainability. Liu et al. (2025) applied ML models to predict rural residential carbon emissions from spatial-form variables, exploring optimization of settlement patterns. While innovative, this approach requires detailed GIS data and may not generalize to urban or non-mapped regions.

C. Interpretability and Explainability

Chen et al. (2025) proposed tree-based models with SHAP and permutation importance to analyze how building characteristics and operation drive emissions. This work emphasizes interpretability but focuses on design/retrofit optimization rather than simple user questionnaires for households.

D. Building Energy Consumption

Davis et al. (2024) compared MLP ANN, SVM, and Decision Tree for non-domestic building energy prediction, finding ANN yielded the lowest MAE/RMSE. Wang et al. (2024) proposed an improved LSTM (ILSTM) for university building energy forecasting in China, achieving lower errors than comparison methods. Zhang et al. (2024) used CNN–LSTM hybrids for short-term load forecasting in commercial buildings, demonstrating significant improvements over traditional LSTM and SVR. However, these studies target specific building types or require high-frequency sensor data, limiting their applicability to simple household-level tools. Li et al. (2024) compared Random Forest, Gradient Boosting, and SVM for building energy records, finding Random Forest provided the best generalization. Their work reinforces the value of tree-based ensembles but does not explicitly predict CO_2 emissions or provide user interfaces.

E. Comprehensive Surveys

Smith and Rodriguez (2025) conducted a comprehensive survey of ML/DL techniques for carbon emission estimation across industrial, transport, building, and national sectors. They highlighted gaps in interpretability and actionable models but did not implement a concrete end-user tool. Bourdeau et al. (2019) provided an earlier review of ANN, SVM, Gaussian processes, and ensembles for building energy forecasting, establishing historical context. Li et al. (2020) summarized traditional and data-driven methods, though their coverage predates recent DL/transformer models and user-oriented tools.

From the literature review, the following research gaps emerge:

(1) Most studies focus on aggregate (national/city) scales or specific building types, not general household-level prediction.

(2) Existing tools often lack interpretability, making it difficult for users to understand emission drivers.

(3) Few works deploy their models in accessible, user-facing applications aligned with SDG-13 goals. This project addresses these gaps by developing a comparative ML framework that predicts household energy and CO_2 emissions, provides interpretable insights, and deploys the model in a simple web dashboard.

III. PROPOSED METHODLOGY

A. Problem Formulation

The goal is to predict household energy consumption (kWh) and associated CO_2 emissions (kg or tons) from easily obtainable building and usage features. Formally,

Given a dataset $\mathbf{D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}}$ where

- $\mathbf{x_i} \in \mathbb{R}^d$ is a feature vector representing building characteristics (floor area, building type, construction material, age) and energy-use attributes (electricity consumption, LPG usage, appliance presence, occupancy, climate indicators), and
- $\mathbf{y_i} \in \mathbb{R}$ is the target variable (energy consumption in kWh or CO_2 emissions in kg)

The objective is to learn a function f(x) such that

$$\hat{y} = f(x) \approx y$$

where ŷ minimizes prediction error measured by Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and maximizes the Coefficient of Determination (R^2).

B. System Architecture

The proposed system follows a four-phase pipeline.

Phase 1 - Data collection and preprocessing:

Identify and integrate open datasets on residential/building energy use and CO_2 emissions (or generate synthetic data if needed). Clean the data (handle missing values, outliers), encode categorical features (one-hot encoding), normalize/standardize numeric variables, and split into training (70%), validation (15%), and test (15%) sets.

Phase 2- model development and comparison.

We implement and compare the following models: (1) **Baseline Models**: Multiple Linear Regression and Polynomial Regression.

(2) **Advanced Ensemble Models:** Random Forest Regressor, Gradient Boosting Regressor, XGBoost Regressor.

(3**) Neural Network:** Small feed-forward MLP with 2–3 hidden layers.

Train each model using k-fold cross-validation (k = 5 or 10) and tune hyperparameters (tree depth, learning rate, number of estimators, MLP hidden units) using grid or random search. Evaluate on the test set using MAE, RMSE, and R^2. Select the best-performing model as the core predictor.

Phase 3 - Interpretability and insights.

Use feature-importance measures (for tree-based models) and SHAP values or permutation importance to identify which factors contribute most to predicted emissions. Generate partial-dependence plots or simple what-if analyses (e.g., reducing AC hours, changing fuel type) to show potential emission reductions.

Phase 4 - Application / Dashboard Development.

Develop a lightweight web interface (Streamlit or Flask) where users enter their building and usage details and receive predicted energy use and CO_2 emissions. Display key drivers and simple recommendations (e.g., "improving insulation" or "switching to efficient appliances") to link predictions with SDG-13 climate-action behavior.

C. Algorithms Used

1. Multiple Linear Regression:

A baseline model that assumes a linear relationship between features and target:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_a x_a.$$

2. Polynomial Regression:

Extends linear regression with polynomial terms to capture non-linear relationships.

3. Random Forest Regressor:

An ensemble of decision trees trained on bootstrap samples, reducing variance and providing feature importance:

$$\hat{y} = (1/T) \sum_{i=1}^{T} h_i(x)$$

where T is the number of trees and h_i is the i-th tree.

4. Gradient Boosting Regressor / XGBoost:

Sequential ensemble methods that build trees iteratively to minimize residual errors. XGBoost includes regularization and parallel processing for improved performance:

$$F(x) = \sum_{m=1}^{M} f_m(x)$$

where f_m is the m-th weak learner.

5. MLP Regressor (Feed-Forward Neural Network):

A small neural network with 2–3 hidden layers using ReLU activation and trained via backpropagation:

$$\hat{y} = \sigma(W_2 \sigma(W_1 x + b_1) + b_2)$$

where W and b are weights and biases, σ is the activation function.

6. Model Evaluation

- Mean Absolute Error (MAE) measures the average absolute difference between predicted and actual values.
- Root Mean Squared Error (RMSE) is the square root of mean squared errors, penalizing large errors.
- R^2 Score represents the proportion of variance explained by the model.
- k-Fold Cross-Validation provides average performance across k splits for robust evaluation.

IV. EXPERIMENTAL SETUP

A. Dataset Description

The dataset is sourced from publicly available residential/building energy datasets (e.g., building energy consumption datasets and residential CO_2 datasets from prior research or open repositories) adapted for this project.

Features (inputs):

- Building characteristics: floor area (m^2), number of rooms, building type (apartment/independent), construction material (concrete/brick/wood), age of building (years).
- Energy-use attributes: annual or monthly electricity consumption (kWh), LPG/other fuel usage (kg or Liters), presence of major appliances (AC, heater, geyser), occupancy patterns (number of occupants, hours at home), climate/temperature indicators (heating/cooling degree days, average temperature).

Target variables (outputs):

- Total energy consumption (kWh or kWh/m^2 per year or month)
- Estimated CO_2 emissions (kg or tons per year, calculated using standard emission factors for electricity and fuels, e.g., 0.82 kg CO_2/kWh for electricity, 2.98 kg CO_2/kg for LPG).

The data consists of structured tabular data stored as CSV files, expected to contain a few thousand records ($n \approx 2000–5000$), suitable for training on a personal PC without GPU.

Preprocessing:

1. Involves handling missing values (imputation or removal)
2. Encoding categorical variables (one-hot encoding for building type, construction material)
3. Feature scaling/normalization (StandardScaler or MinMaxScaler)
4. Train-validation-test split (70-15-15).

B. Implementation Details

The implementation uses Python 3.10+ (Anaconda), Jupyter Notebook or VS Code, and Git for version control.

Libraries and packages include:

1. **Data handling:** pandas and numpy.
2. **ML Models:** scikit-learn (LinearRegression, RandomForestRegressor, GradientBoostingRegressor, train_test_split, metrics), XGBoost, and sklearn.neural_network (MLP Regressor).
3. **Visualization:** Matplotlib and seaborn.
4. **Explainability:** Shap and sklearn permutation importance
5. **Web application:** Streamlit or Flask for the.

The implementation workflow consists of five steps:

(1) Load CSV datasets, perform cleaning, encoding, normalization, and split.

(2) Train baseline models (Linear/Polynomial Regression), then ensemble models (Random Forest, Gradient Boosting, XGBoost) and MLP using k-fold cross-validation and hyperparameter tuning.

(3) Evaluate each model with MAE, RMSE, and R^2, select best model and generate plots (predicted vs. actual, error histograms).

(4) Compute feature importance/SHAP values, derive insights on top drivers.

(5) Wrap selected model and preprocessing pipeline using joblib/pickle and integrate into Streamlit app.

V. PRELIMINARY RESULTS AND DISCUSSION

Based on the literature review, we anticipate the following expected outcomes.

1. **Baseline Performance:** Linear and Polynomial Regression are expected to achieve moderate R^2 (0.60–0.75) and serve as benchmarks.
2. **Ensemble Model Performance:** Random Forest, Gradient Boosting, and XGBoost are expected to outperform baselines, achieving R^2 greater than 0.85, MAE less than 100 kg CO_2/year (or equivalent energy units), and RMSE less than 150 kg CO_2/year,

consistent with Khan et al. (2025) and Li et al. (2024) .

3. **Neural Network Performance:** The small MLP is expected to perform comparably to tree ensembles but may require more tuning. It serves as a lightweight deep-learning baseline.
4. **Feature Importance:** Preliminary analysis suggests floor area, electricity consumption, and appliance presence (AC, heater) will be the top predictors of emissions, aligning with Chen et al. (2025) .
5. **Interpretability:** SHAP values will provide clear visualizations of how each feature contributes to predictions, enabling actionable insights for users. Once actual experiments are complete, this section will include tables and graphs showing MAE, RMSE, R^2 for each model, feature importance plots, and predicted vs. actual scatter plots.

VI. CONCLUSION AND FUTURE WORK

This project proposes a machine learning framework to predict household energy consumption and CO_2 emissions from easily collected features, addressing SDG-13: Climate Action. Preliminary design indicates that tree-based ensemble models (Random Forest, XGBoost) will provide accurate, interpretable predictions suitable for deployment in a user-facing web dashboard.

Key contributions include:

(1) Comparative evaluation of multiple ML models for household-level carbon/energy prediction.
(2) Interpretable outputs via feature importance and SHAP analysis.
(3) Practical deployment as a web tool for personalized emission reduction guidance.

Future work will focus on:

(1) **Richer and Localized Datasets:** Extend with locally collected data (Indian households, campus buildings) for regional relevance.
(2) **Advanced Time-Series Models:** Explore LSTM or CNN-LSTM for finer-grained (hourly/daily) forecasts.
(3) **Demand-Side Management:** Integrate optimization modules for appliance scheduling to minimize emissions.
(4) **User Studies:** Conduct studies to measure dashboard impact on awareness and behavior change.
(5) **Real-Time Integration:** Connect to IoT smart-meter data or public APIs for near real-time tracking.

References

[1] X. Ye et al., "An examination of daily CO_2 emissions prediction through machine learning and deep learning models," *Environmental Research Letters*, Nature Portfolio, 2025.

[2] J. Smith and L. Rodriguez, "Machine learning applications for carbon emission estimation," *Cleaner Engineering and Technology*, Elsevier, 2025.

[3] A. Khan et al., "Predicting residential CO_2 emissions: Machine learning and deep learning approaches," *Sustainability in Building and the Environment*, Durabi, 2025.

[4] H. Liu et al., "Machine learning models for predicting rural residential carbon emissions and optimising spatial forms," *Scientific Reports*, Nature, 2025.

[5] L. Chen et al., "An interpretable machine learning framework for residential building energy and carbon emissions," *Scientific Reports*, Nature, 2025.

[6] M. Davis et al., "Building energy consumption prediction using deep learning," *Applied Energy*, 2024.

[7] H. Wang et al., "Intelligent prediction method of building energy consumption based on deep learning," *International Journal of Photoenergy*, Hindawi, 2024.

[8] K. Zhang et al., "A deep learning framework for building energy consumption forecasting," *Renewable & Sustainable Energy Reviews*, 2024.

[9] F. Li et al., "Machine learning algorithms for predicting energy consumption in buildings," *Energy Reports*, 2024.

[10] R. Ahmed et al., "Forecasting CO_2 emission in the US using machine learning methods," *Journal of Data Science and Intelligent Systems*, 2023.

[11] J. White and P. Jones, "Building energy prediction models and related uncertainties," *Buildings*, MDPI, 2022.

[12] M. Shapi et al., "Energy consumption prediction by using machine learning," *Energy Reports*, Elsevier, 2021.

[13] X. Li et al., "A state-of-the-art review on the prediction of building energy consumption," *Building Services Engineering Research & Technology*, 2020.

[14] M. Bourdeau, X. Zhai, E. Nefzaoui, X. Guo, and P. Chatellier, "Modeling and forecasting building energy consumption: A review of data-driven techniques," *Sustainable Cities and Society*, vol. 48, p. 101533, 2019.

[15] V. Madhusudanan, "A machine learning framework for energy consumption in buildings," Master's thesis, Clemson University, 2019.

[16] A. Almalaq and G. Edwards, "Evolutionary deep learning-based energy consumption prediction for buildings," in *IEEE* proceedings, 2018.

[17] M. Rahman and S. Ali, "Application of deep learning model in building energy consumption prediction," *Computational Intelligence and Neuroscience*, Hindawi, 2016.