Improving Residential Energy Efficiency Using Machine Learning: A Predictive Approach to Smart Home Energy Optimization

MSc Research Project
FinTech

Mohammed Aslam Nalakath Assi
Student ID: x22207708

School of Computing
National College of Ireland

Supervisor: Faithful Onwuegbuche

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Mohammed Aslam Nalakath Assi
Student ID:	x22207708
Programme:	FinTech
Year:	2025
Module:	MSc Research Project
Supervisor:	Faithful Onwuegbuche
Submission Due Date:	11/08/2025
Project Title:	Improving Residential Energy Efficiency Using Machine Learning: A Predictive Approach to Smart Home Energy Optimization
Word Count:	6768
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Mohammed Aslam Nalakath Assi
Date:	13th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	■
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	■
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	■

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Improving Residential Energy Efficiency Using Machine Learning: A Predictive Approach to Smart Home Energy Optimization

Mohammed Aslam Nalakath Assi
x22207708

Abstract

Residential buildings are a large contributor of energy consumption and emissions, thus efficiency improvements are a milestone. The paper identifies and develops implications of interpretable machine learning, classifying and predicting residential energy efficiency based on the U.S. dataset of NREL ResStock. Aspects included structural as well as insulation, mechanical as well as socioeconomic features. The data preprocessing involved the removal of columns, imputing missing values, encoding categorical data and scaling the features, where labels were Efficient, Moderate or Inefficient. The 80/20 split was used to test three classification and three regression models. Random Forest models produced the best results, 87% accuracy (F1 = 0.88), and 0.94 R^2 outperforming other comparable research. SHAP analysis showed that the key predictors were CO_2 emissions, square footage, and the type of the heating and cooling system, allowing guided recommendations on retrofits. The findings indicate that balanced datasets, variety of features, and ensembles achieves accurate, transparent energy assessment prediction models. Limitations involve cybernetics dependency on simulated data and lack of the complete retrofit module which is found to be improved in coming work at combination with model optimisation and adjustment to diverse housing stocks.

1 Introduction

1.1 Background and Motivation

The residential buildings consume large percentages in the world energy demand which is estimated to be about 30% of final energy as well as electricity being about 35% of the same share (International Energy Agency, 2022). Not only is this demand one of the biggest contributors to emissions of greenhouse gases, but it also causes households to incur financial expenses, particularly those of the low-income and middle-income groups. Energy efficiency of residential buildings have become an environmental and social priority as countries pledge to move towards net-zero carbon targets and energy sustainability.

Improving residential energy performance is a complicated issue, even though it is of great significance. Conventional energy audits and rule-based simulation models (including ones based on building codes or engineering assumptions) tend to be labour-intensive and typically cannot be easily scaled to the national or regional level. Also, they might not sufficiently reflect field based variations in behaviour and contextual differences in various homes. Data-driven regression-based approaches for residential energy consumption modelling have been shown earlier to hold strong potential in providing scalability and economic comparability compared to more traditional engineering-driven techniques(Fumo and Biswas, 2015).

1.2 Importance of the Research

The study is based on the premise that the majority of the past energy efficiency research places its main emphasis on physical or structural characteristics of houses leaving behind behavioural, climatic and socioeconomic factors, which play a huge role in the consumption of the energy. What is more, although black-box ML models can be rather accurate, they are not always transparent, which restricts their application in the real world including policy planning, household consultation, or retrofit programs.

This paper will solve such issues by building explainable ML models that:

- Include numerous functions, such as insulation, use of appliances, heating/cooling systems, level of income, and geographic zone
- Predict energy consumption per square foot (a normalized efficiency metric) and classify homes into *Efficient*, *Moderate*, or *Inefficient* categories
- Use SHAP (SHapley Additive exPlanations) to show contribution of every feature to the final prediction

By so doing, the study also makes a contribution to the efforts of global sustainability as it aids in two reasons of Global Goals by the United Nations Sustainable Development Goals (SDGs):

- **SDG 7: Affordable and Clean Energy** - this goal should be promoted by making accessible, data-based assessment tools that can be used in decreasing household energy burdens
- **SDG 13: Climate Action** — With the support of the provision of essential retrofits and improved awareness of energy use on the household level, it will be possible to reduce the emission by implementing the necessary actions to achieve this goal

1.3 Research Question and Objectives

Research Question:

How can interpretable machine learning models be used to predict and classify household energy efficiency in order to support data-driven retrofit strategies and contribute to environmental sustainability?

Research Objectives:

- To pre-process a complex dataset of residential energy with behaviour, environment and structural variables

- To normalize energy consumption variable (`energy_per_sqft`) and deduce a label to show efficiency levels of multiclass
- To build and analyze machine learning models in order to perform both classification and regression tasks
- To apply SHAP explainability methods to explain model output and identifying some main drivers of performance
- To put the model to test on real world inputs and simulate retrofit prescriptions on inefficient home

These goals are meant to fill the trade-off between accuracy of prediction and explainability, which can assist in substantive, practical insights at the household level.

1.4 Scope and Limitations

The study is subject to the assumption that the data and the sample represent a general residential population of the U.S. and that all the recorded information (e.g., the insulation factor, HVAC structure type, the use of appliances, etc.) is correct. It is taken on pseudo-static data, not taking into account real-time sensor data, dynamically varying occupancy schedules, or energy prices. This way, the findings capture a general trend as opposed to operational behaviours of a household.

1.5 Structure of the Report

This report is structured as follows:

- **Chapter 2 – Literature Review:** Examine available scholarly and business studies in the field of residential energy efficiency, predictive modelling and applications of predictive modelling in energy and sustainability sector.
- **Chapter 3 – Research Methodology, Design and Implementation:** Gives an account of the methodology involved in the preprocessing of the dataset, feature engineering, and energy efficiency categories, labeling, and machine learning models (classification and regression). The technical design and the implementation workflow are presented.
- **Chapter 4 – Results:** Presents the results of the experiment, i.e., the performance of the models, SHAP model interpretability analysis, and the comparison of models. Predictive performance and relevance of features are visualized in both tables and visualizations.
- **Chapter 5 – Discussion and Conclusion:** Concludes with the summary of the most important points, discusses the implication of the findings to its application in practice, outlines limitations and provides future research directions

2 Related Work

The residential sector is a large consumer of the global energy supply and emits significant quantities of greenhouse gases; hence, data driven efficiencies will be an effective lever towards achieving SDG 7 (Affordable and Clean Energy) and SDG 13 (Climate Action). Household heating and efficiency are put at the centre of decarbonisation routes by the International Energy Agency (2022), whilst ground breaking work in building energy modelling indicates the usefulness of data centric methodologies (Fumo and Biswas, 2015). The recent bibliometric and systematic reviews verify the fact that the methods of machine learning (ML) and deep learning (DL)-based building energy optimisation experience a sharp increase (Liu and Chen, 2025; Husseini et al., 2025; Aghili et al., 2025). They use more detailed data, such as smart meters, building characteristics, weather, occupancy, and socioeconomic signals, but at scale, there are challenges in regards to interpretability, transferability, data paucity, and the ability to convert predictions to retrofit/control choices (Rehman et al., 2025; Raza et al., 2024). The remainder of this chapter brings a synthesis of predictive modelling, explainable AI, behaviour and occupancy modelling, advanced DL, energy management systems and policy/scale considerations and sums up identifying the gaps this report will fill.

2.1 Predictive Modelling for Residential Energy

One of the central streams of works refers to supervised ML of predictive residential energy consumption and loads. Asamoah and Shittu (2025) compare HVAC load prediction in residential buildings by gauging the potential of Gradient Boosting and Random Forest models and demonstrates that these methods are superior to linear baselines, yet perform poorly during extreme weather events- an indication that models must be able to perform well in situations where climatic conditions are highly variable. Cui et al. (2024) build predictive models across the different types of residential buildings and demonstrate based on the post hoc analysis that the envelope quality and the type of HVAC system play a dominant role in Omega variation. Parallel to this, Mahmood and Asif (2024) compare supervised algorithms to predict residential energy efficiency, and conclude that envelope and system properties are always among the best predictors.

On larger levels, Peplinski et al. (2024) combine smart meter, building, climate, and socioeconomic data in order to predict residential electricity demand; their study also points to the strengths of multi source merging and cautious feature engineering. Though the research was done in an office setting within a university building, Yesilyurt et al. (2024) evaluate the multiple algorithms on building loadings and the influence of the operational schedules, which can be applied to residential areas. Uddin et al. (2024) reconcile the spaces between spatial design and prediction by integrating agent based simulation and ML in measuring the interaction between the indoor layout configuration and occupant behaviour in determining consumption.

Some of the works promote higher capacity or hybrid versions when they are comparing them with the lighter baseline. Lotfipoor et al. (2024) combine the empirical mode decomposition and Bayesian optimisation with deep neural networks to enhance residential load forecasting and reveal superior benefits compared to baseline in particular to longer horizons. In Bozorgi et al. (2025) model multi unit building cooling using a hybrid ML DL controller, energy efficiency is enhanced through a mere consideration of the computational-time. In addition to these, Zoubir et al. (2024) show building energy consumption predictive modelling based on traditional ML on mixed data, with an emphasis on the reliability of the data and features coverage.

2.2 Explainable AI and Interpretability in Energy Modelling

Interpretability relates prediction with action. Cui et al. (2024) couple ML with SHAP to reveal the profoundness of insulation and the potential HVAC type into turning feature rankings into the priorities of retrofitting. In order to handle data scarcity, Rempi et al. (2025) use CTGAN to create synthetic samples; they subsequently use SHAP to interpret the classification of retrofit, allowing analysis in under sampled areas but warning of the potential biases of synthetic data. On the operations front, Nigar et al. (2024) develop an Energy Advisor used in the demand side management where the behavioural and appliance level interventions are informed through SHAP style feature rankings. Ali et al. (2024) can combine ML with GIS at city scale to map retrofit potential and evaluate performance and utilise interpretability to preserve transparency with stakeholders. Summarising the findings of these broad review papers, a certain lesson emerges: explainable approaches are needed to build trust and readiness and, yet, underutilised in production pipelines (Aghili et al., 2025; Husseini et al., 2025). As echoed, the systematic review of smart home energy management by Rehman et al. (2025) presented interpretable models as closest to both user acceptance and the needs of policy. .

2.3 Occupancy, Behaviour, and Socioeconomic Factors

Occupancy and behaviour influence household energy consumption in a material way. Incredibly, Lee et al. (2022) demonstrate that adaptive HVAC control and load reductions become unlockable because of DL-based occupancy detecting based on appliance usage. The above is confirmed by Uddin et al. (2024), who demonstrate that, in addition to the living habits, the indoor layout and movement patterns moderates the demand, hence the interactions between occupants and spaces ought to be modelled, rather than be considered as noise. Substantial load shifts due to occupancy schedules in office buildings were observed and documented by Yesilyurt et al. (2024), which is an analogue to multi-family residential buildings whose occupancy schedule and density varies. It depends on the socioeconomic setting as well. Peplinski et al. (2024)indicate that by including social-economic variables to meter and weather data, one is able to predict and segment better. At the appliance level of behaviour, Steephen et al. (2022) survey deep learning to non intrusive load monitoring (NILM) demonstrating that personalised feedback and conservation are possible due to appliance signatures in methods, yet identify privacy and data collection limitations. Roach (2020) supplements this description with a mixed model analysis that connects building attributes to electricity profiles, and Qiao and Yunusa-Kaltungo (2023) introduce a hybrid agent based method of ML that takes human behaviour as the focus of energy forecasts.

2.4 Advanced Modelling Approaches: Deep Learning

DL has worked well on non linear data on high dimensions energy data, especially when rich temporal signal is present. It is shown by Lee et al. (2022) that occupancy dependent appliance trends can be correlated to DL to achieve real time control. As demonstrated by Lotfipoor et al. (2024), LSTM style networks optimised using Bayesian and assisted by EMD generate good forecasts on the residential load, particularly at longer horizons. Mehmood et al. (2023) apply ANNs and then add weather metrics weighting to enhance smart home prediction over exogenous variability.

In application to systems, Bozorgi et al. (2025) use a hybrid CNN LSTM block in a cooling control framework of a multi unit residence where a higher level of accuracy is exchanged with compute. Khan et al. (2020) offer a combination of CNN + LSTM Autoencoder to residential and commercial forecasting that attains reduced errors when compared to single model baseline but with increased training complexity. At community scale, Moosbrugger et al. (2025) compared DL (LSTM/Transformer) and simpler baselines in short term forecasting, finding that in data scarce conditions, lean ML or persistence can lead to competitive results–a plus again in favour of fit to purpose modelling. Two general surveys, done by Husseini et al. (2025) and Raza et al. (2024), come to the conclusion that, despite achieving a state of the art accuracy in forecasting, NILM, and control, the uptake of DL is tempered by computational requirements, data needs, and the absence of native understandability.

2.5 Energy Management Systems and Demand Side Optimisation

To translate the predictions to savings necessitates the incorporations with energy management systems (EMS) and the demand side management (DSM). Nigar et al. (2024) offer an ML-enabled Energy Advisor that can restructure load profiles over tariffs and peaks, explaining which behavioural/appliance actions should be prioritised to consume energy. The survey by Raza et al. (2024) surveys smart home EMS designs, where the ML/DL enabled scheduling, load shifting, and multi objective control is mapped, whereas the survey by Rehman et al. (2025) stresses closer connections between forecasting, automated control, and user feedback loops. On planning scale, Ali et al. (2024) demonstrate how the ML + GIS will define the city scale retrofit mapping and investment prioritisation. The study by Powroźnik and Szcześniak (2024) shows that predictive analytics has the potential to minimize the peak demand with possible comfort constraints. Mehmood et al. (2023) integrate weather and occupancy adaptive setpoints via predictive models in smart home controllers; Nandkeolyar et al. (2025) take this further with adaptive DSM and weather based occupancy due to battery storage with distributive generators, demonstrated in a hybrid model of DNNs–probably another opportunity to provide feeder balance via control oriented ML.

2.6 Policy, Climate Action, and Large Scale Implementation

To scale data driven efficiency, there is the need of policy and solid body of evidence. Ali et al. (2024) introduce the relation between retrofit planning to the urban climate action allowing the cities the possibility to measure their contribution to SDGs. According to Rehman et al. (2025), governance and incentive are vital in the adoption of EMS and real time control. The works by Aghili et al. (2025) and Husseini et al. (2025)are based on the idea that low overhead and transparent models are more appropriate as models to deploy in the public sector and gain trust and enjoy use by occupants. Liu and Chen (2025) report the field development and demand collective datasets and benchmarks in enhancing comparability and transferability. The evidence base is rounded out by neighboring strands. Sherif et al. (2025) incorporates microclimate and socioeconomic data into localised UBEM to increase spatial granularity; Thorve et al. (2023) publishes high-resolution synthetic residential profiles to facilitate large scale modelling; Balti et al. (2024) incorporates occupancy detection to better disaggregate energy, and downstream manage energy; Alzoubi and Mishra (2024) argue that Green AI trade offs tend toward lightweight AI with explainable solutions; Gorzałczany and Rudziński (2024) demonstrate how fuzzy and evolutionary methods can help to achieve the balance between low accuracy and transparency; Dell'Anna (2025) relates ML based assessments of EPC effectiveness to real estate outcomes- useful in designing incentives and regulatory levers.

2.7 Research Gaps

Although recent developments have achieved progress in energy efficiency analytics, it left significant gaps. There is not much integration between regression models (e.g., kWh per household or per square foot) and classification based efficiency labels. These, combined, would enhance prioritisation and lead to more intuitive outputs in terms of both output to households and to policymakers. Explainability remains primarily limited to feature ranking, and makes little use of formal rule-outs across structured domains that can guide action directly.

Big solutions also dominate the field with deep learning models as it is shown that small, scalable models that can work with missing data and drift, and that perform under privacy constraints are underrepresented. In data sparse areas, synthetic data (e.g., CTGAN Rempi et al. (2025)) are promising but must undergo rigorous bias auditing and need to be used judiciously lest they perpetuate biases. Lastly, clear matching to SDGs and utilisation of standardised impact measures would facilitate ease of comparison, uptake, and funding of results to facilitate general adoption of policy.

2.8 Summary

Literature reflects that ML/DL has been able to accurately predict residential energy consumption; explainable AI has the potential to unveil driver actions; and EMS/DSM experiments have found that model integration into control and planning processes can provide operational improvement. But the bridge to predictions can be trusted, scaled to solve, and connected with SDGs is not there yet. This dissertation fills these gaps by integrating consumption prediction with multiclass efficiency labelling, with SHAP directed recommendations to propose retrofit and behavioural measures and focuses on lightweight, explainable models in a manner that suits real world applications.

3 Methodology

3.1 Research Design Overview

The research design utilized in this study is a quantitative, experimental research design whereby machine learning will be applied to the prediction and classification of the residential energy efficiency. The method implies the process of analytically processing data and analysing it in order to find out some relevant patterns, model predictive models, and evaluate model performance.

The study would focus on accuracy as well as interpretation such that other than the findings being accurate, they are interpretable and can be applied in practical situations in the making of decisions. The paper proposes to harmonize predictive analytics and explanatory analysis with the objective to facilitate a decision-making action plan to be more likely to increase residential energy efficiency.

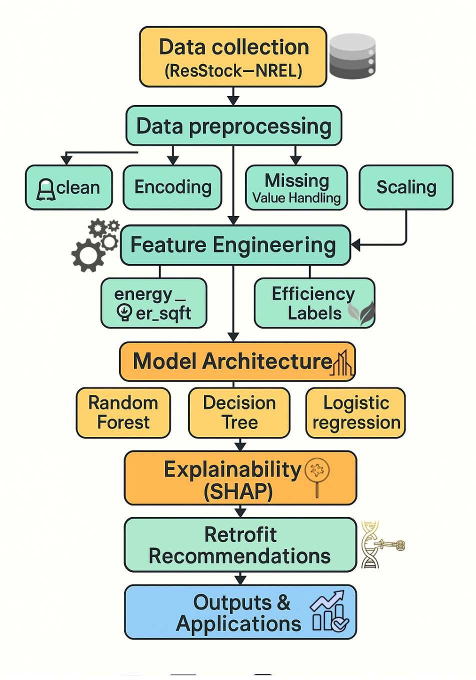

Figure 1: End-to-end pipeline for residential energy efficiency modelling.

3.2 Data Source

The data in this research was taken under the ResStock high resolution residential building stock model which is set up and updated at the National Renewable Energy Laboratory (NREL)[1]. ResStock is a mash-up of the data provided by various national surveys and field studies, geospatial data to create a statistically representative sample of residential buildings nation wide within the United States. Within each record, large variety of attributes is provided that defines physical, mechanical and socio-economic nature of the property.

The structural features contained in the dataset comprise floor area (in square feet), amount of beds, insulation of walls, attics, and floors, and the amount and the kind of windows. Some of the information available on system includes the main HVAC system type (e.g. central air conditioning, heat pump, furnace) and the type of fuel used to heat the water (e.g. natural gas, electricity, propane). The socioeconomic indicators are household income, percentage of area median income (AMI), representative household income brackets.

[1] `https://resstock.nrel.gov/datasets`

3.3 Data Preparation and Feature Engineering

3.3.1 Column Removal

The topics of irrelevant and non-informative variables were handled systematically in the preprocessing step, which excluded unique building identifiers, dataset indexing fields and metadata attribute which have no direct or indirect effect on energy performance. The action was taken to reduce noise in the data, decrease computational burden and avoid overfitting that may result due to non-predictive features in the model.

3.3.2 Handling Missing Data

The missing values were handled in a well organized type-specific way. First, features were divided into different categories based on data type: columns with a float64 and int64 data-type were trained as numeric, object and category columns were trained as categorical.

Missing values were imputed in numeric features based on median values of column (which were used to reduce the effect of outliers and skewed patterns). In case of the category parameter, any missing values were imputed using the mode which kept the most frequently observed level in that variable.

Moreover, those features that had a high percentage of missing data (above the configured threshold, e.g., 80%) were removed as a whole to ensure good quality of the dataset and keep model training stagnant. That was done programmatically in Python, which would make such processing consistent and reproducible throughout the dataset.

3.3.3 Categorical Encoding

The categorical feature such as insulation level and type of HVAC were converted into numbers so that they could be compatible with machine learning algorithms. Low-cardinality variables were one-hot encoded, with the result that a binary feature column was produced per category with no ordinal relationship enforced. This avoids the misinterpretation of the algorithms that a nominal categories do have an original order.

With respect to ordinal variables, whose main difference was ranked in a clear manner, the assumption of label encoding was implemented to give the integer values that reflect their natural order. By doing this, the relative level in the categories is maintained and model input is efficient. Encoding was done in a code-based manner to allow consistency and repeatability throughout the dataset.

3.3.4 Feature Scaling

For algorithms sensitive to feature magnitude, numerical variables were standardized using the `StandardScaler` from `scikit-learn`, transforming each feature to have zero mean and unit variance:

$$x' = \frac{x - \mu}{\sigma}$$

This ensured that all features contributed equally to model training and prevented scale dominance.

3.3.5 Target Construction

Energy per square foot (major target variable) was calculated by dividing the property characters Total Annual Energy Consumption (kilowatt-hours) with the Floor Area (square feet) of every property:

$$\text{energy\_per\_sqft} = \frac{\text{Total Annual Energy Consumption (kWh)}}{\text{Floor Area (sqft)}}$$

This normalisation takes into account the difference in building size so that comparisons between buildings of differing sizes could be made on equal terms. The measure is better suited to predictive model building and classification activities by offering a standardised measure of energy efficiency comparisons that can remove the influence of building area on a performance outcome

3.3.6 Label Engineering

A second target variable, `efficiency_label`, was defined as a categorical variable defined by relating properties to efficiency classes characterised by specified energy-per-square-foot thresholds:

- *Efficient*: $\leq$ 8.0 kWh/sqft
- *Moderate*: 8.01–12.0 kWh/sqft
- *Inefficient*: $>$ 12.0 kWh/sqft

These wiring were chosen in a way that they contained significant differences in residential energy performance. The derived categorical variable can be used to estimate multi-class classification models as well as regressions giving two complementary viewpoints modeling how to predict the level of consumption exactly and classify properties into levels of efficiency to target the decision making and policy making.

3.4 Modelling Tasks and Evaluation Plan

3.4.1 Regression Models

Regression models were created that estimate the residential energy efficiency across the properties of varying size and characteristics in the interest of accurately estimating the values of the continuous `energy_per_sqft` target variable. There were three algorithms used:

- Linear Regression : assumption of a proportional relationship between predictor and the target variable was used as a basis, whereas the outcome variable was provided in a linear, directly proportional manner. It can be used to set a benchmark that the other, more complex models can be compared against since it is transparent and interpretable.
- Decision Tree Regressor : A non-parametric model that can model non-linear relationships and the effect of interactions among the features without emphasis on features scaling. It makes it especially suitable on heterogeneously typed data.

- Random Forest Regressor : An ensemble that averages several heavily-weighted decision trees in a training set based on the bagging method to create more stable and better predictions as well as minimize the chance of overfitting..

Three statistics that are common among regression models were used to measure model performance:

- Mean Absolute Error (MAE): Measures the average size of the prediction errors in the units in which the target is measured, such that an interpretation of accuracy is easily provide
- Mean Squared Error (MSE): Like MAE but squares the errors, so large deviations are punished more and the large prediction discrepancies are emphasised
- R-squared (R^2): R-squared- shows the percentage of variance in the outcome variable to be explained in by the model, as a primary indicator of all-over model fit.

3.4.2 Classification Models

Classification models were created in order to forecast the categorical target variable efficiency_label in which each property is assigned into one of three classes: (*Efficient*, *Moderate*, *Inefficient*)The three algorithms were adopted:

- Logistic Regression : it was used as a baseline probabilistic model generating probabilities of classes with the use of a linear decision boundary. It gives an easy referencing scale to more advanced models.
- Decision Tree Classifier : A rule-based model that divides data into decision nodes using the values of the different features, enabling it to represent non-linear relationships, as well as the interaction between features.
- Random Forest Classifier : An implementation of multiple decision trees whose predictions are combined to increase accuracy, generalisation and minimise overfitting.

The standard classification metrics were used to evaluate model performance:

- Accuracy : A ratio of good samples recognized scrupulously as compared to all groups.
- Precision, Recall, F1-score : Gives a balanced impression of model performance, which is particularly relevant when the distribution of the classes is not balanced, where precision addresses the rate of correct positive predictions, recall the proportion of positives being assigned a positive prediction and F1-score is the harmonic mean thereof.
- Confusion Matrix : Provides a break down analysis of the right and wrong prediction for each of the classes allowing one to identify individual areas where the model works well or are problematic.

3.5 Explainability

SHAP (SHapley Additive Explanations) was used to interpret the model in the way it allowed predicting the features that would need improvement in terms of interpretation. Internally, dietary calcium, the most important variables of the model was ascertained at the global level using SHAP ranking of features by their average contribution to the model output. It illustrated predictions at an individual level, demonstrating the influence on the result to each household based on specific values of the features. These insights were visualised in a understandable way through summary plots, force plots and decision plots that made the models to be accurate and transparent.

3.6 Retrofit Simulation

Retrofit simulation module will also be used to determine the efficiency increases in homes that were assessed as*Inefficient*. These include making programmatic adjustments on the chosen characteristic (e.g. level of insulation, type of HVAC system) to reflect more realistic upgrade scenarios and running predictions again to assess what effect this has on `energy_per_sqft` and efficiency classification. Although the basis of this module has been developed, evaluation and testing is being undertaken.

4 Design Specification

4.1 System Architecture

The proposed system for predicting and classifying residential energy efficiency is organized as a modular pipeline:

1. **Data Ingestion**: NREL ResStock dataset with structural, system, climate, and socioeconomic attributes.

2. **Preprocessing Module**: column pruning, missing-value handling, categorical encoding, and feature scaling (for magnitude-sensitive algorithms).

3. **Feature Engineering Module**: generation of engineered variables, including `energy_per_sqft` and `efficiency_label`.

4. **Model Development**: training pipelines for regression and classification families (linear models, decision trees, ensembles).

5. **Evaluation Module**: computation of task-appropriate metrics and storage of structured reports.

6. **Explainability Module**: SHAP-based global and local attributions with standardized visualizations.

7. **Retrofit Simulation Module (planned)**: controlled feature adjustments for homes labeled *Inefficient* and re-evaluation of model outputs.

4.2 Data Schema and Feature Definitions

Core variables used by the system are summarized below.[2]

Table 1: Feature dictionary.

Feature	Type	Category	Description
in.sqft	Numeric	Structural	Heated/cooled floor area (sqft)
in.bedrooms	Numeric	Structural	Number of bedrooms
in.occupants	Numeric	Occupancy	Number of occupants
in.windows	Categorical	Envelope	Window count/level
in.insulation_ceiling	Categorical	Envelope	Ceiling insulation category
in.insulation_floor	Categorical	Envelope	Floor insulation category
in.insulation_wall	Categorical	Envelope	Wall insulation category
in.hvac_heating_type	Categorical	Systems	Heating system type
in.hvac_cooling_type	Categorical	Systems	Cooling system type
in.water_heater_fuel	Categorical	Systems	Water heater fuel type
in.income	Numeric	Socioeconomic	Reported household income
in.representative_income	Numeric	Socioeconomic	Representative income estimate
in.area_median_income	Numeric	Socioeconomic	Median income for the area
total.energy_kwh	Numeric	Target/Input	Annual energy consumption (kWh)
energy.per.sqft	Numeric	Engineered	Normalized annual energy use (kWh/sqft)
efficiency_label	Categorical	Engineered	Efficient / Moderate / Inefficient

4.3 Experimental Design Parameters

The dataset was divided into an 80/20 train–test split using the `train_test_split` function, with `stratify=y` applied for classification tasks to preserve the original class distribution. A fixed random seed (`42`) was used to ensure reproducibility of the splits. To prevent data leakage, all preprocessing steps, including encoding and scaling, were fitted exclusively on the training data and then applied to the test data. Categorical features were encoded according to their nature: one-hot encoding was used for nominal variables, while ordinal variables were label encoded to retain their inherent ordering. Class balance was monitored by inspecting counts and confusion matrices, and no resampling methods were applied in this implementation.

4.4 Model Configuration Overview

Three regression models were developed for predicting normalized energy consumption per square foot: Linear Regression, Decision Tree Regressor, and Random Forest Regressor. For classification tasks, Logistic Regression, Decision Tree Classifier, and Random Forest Classifier were employed. Numerical features were standardized using Stand-

[2]Feature names should match the final dataset exactly for reproducibility.

ardScaler for models sensitive to feature magnitude, such as Linear and Logistic Regression, while tree-based models were trained without scaling requirements but still followed a consistent preprocessing workflow for reproducibility. All categorical variables were encoded as described in Section 4.4 to ensure compatibility with the algorithms. Model parameters used in this study were based on the default configurations provided by their respective libraries, with no additional hyperparameter tuning performed.

4.5 Evaluation Framework

Regression model performance was evaluated on the held-out 20% test set using Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination (R^2). Classification models were assessed on the same test split using Accuracy, Precision, Recall, and F1-score, with results reported both per class and as macro-averaged scores. Confusion matrices were also generated to visualize the distribution of correct and incorrect predictions across classes. To improve interpretability, SHAP (SHapley Additive Explanations) was applied to both regression and classification models, producing summary plots for global feature importance and local explanation plots for selected instances. All reported metrics and interpretations are based solely on the performance of models on the test set.

5 Implementation

5.1 Development Environment

The implementation of this research was conducted entirely in the Google Colab environment using Python 3.10. Colab was selected for its accessibility, compatibility with widely used data science libraries, and integration with cloud storage, which enabled seamless dataset loading and the saving of intermediate artifacts. All computations were executed on Colab's hosted CPU runtime, which was sufficient for the dataset size and model complexity applied in this study.

5.2 Programming Libraries

These are Python libraries that aided data management, model building, superiority and visualization:

- **pandas** and **numpy**: to load, clean and transform a dataset and to perform numerical computation.
- **scikit-learn**: to deploy regression and classification models, preprocessing Train-test splitting, performance metric calculation and the encoding, scaling process.
- **matplotlib** and **seaborn**: when you need to create exploratory data analysis (EDA) plots. confusion matrices, visualizations of performance.
- **shap**: to create outputs of global as well as local interpreting models.

5.3 Implementation Workflow

The workflow was started by the importing of the dataset at the NREL ResStock repository and loading it into a pandas DataFrame. Subsequently preprocessing functions were performed, such as the removal of unnecessary columns, treating the missing values, encoding categorical characteristics, and scaling of numeric variables through StandardScaler to models sensitive to the level of characteristics. Then feature engineering was used to calculate the target variable: energy_per_sqft and label the efficiency with pre-determined labels. There were two distinct target variables: the continuous energy_per_sqft to be used in regressions and the categorical efficiency_label to be used in classification. Each modeling task used 80 percent of the data as training data and 20 percent as testing data, and stratification of the concept labels allowed preservation of the amount of classes. The training data set were fit to preprocessing steps only and were used to compute steps on the test set to prevent the data leakage.

5.4 Model Development

Three regression models—Linear Regression, Decision Tree Regressor, and Random Forest Regressor were implemented to predict `energy_per_sqft`. For classification, Logistic Regression, Decision Tree Classifier, and Random Forest Classifier were developed to assign efficiency labels. Models were initialized with default hyperparameters from scikit-learn, and no additional tuning was performed in the current implementation.

5.5 Evaluation and Explainability

Model performance was assessed using the metrics outlined in the Design Specification: MAE, MSE, and R^2 for regression; Accuracy, Precision, Recall, F1-score, and Confusion Matrix for classification. All metrics were computed on the held-out test set. SHAP was employed to analyze feature importance, generating summary plots for global interpretation and force plots for local explanations. This approach enabled insight into both the overall model behavior and specific predictions.

5.6 Outputs and Artifacts

The deployment generated a completely preprocessed dataset prepared to use in modelling, trained regression and classification models that could be used repeatedly, performance measure tables and visualisations on each model, SHAP importance plots as well as local interpretability plots, and an initial workflow of the proposed simulation module on the retrofit.

6 Results

The following part gives the results of the classification and regression models of the residential energy efficiency prediction with the help of SHAP-based interpretability analysis. The findings are also combined with a comparison of the literature, which makes it possible to position the results of this study against the existing methods.

The overview of the data set during its first inspection showed a significant skew on the labels of efficiency because the Inefficient category made a large share of the dataset.

The bias that arises with class imbalance is that the machine learning model predicts the majority class, which causes a drop in the generalization performance. In response to this, a balancing process was used thereby ending up with the exact proportion of all three categories of efficiency:*Efficient*, *Moderate*, and *Inefficient*.Similar pre-processing techniques have demonstrated that it is possible to enhance the recall of minority classes in studies of residential energy (Mahmood and Asif, 2024)

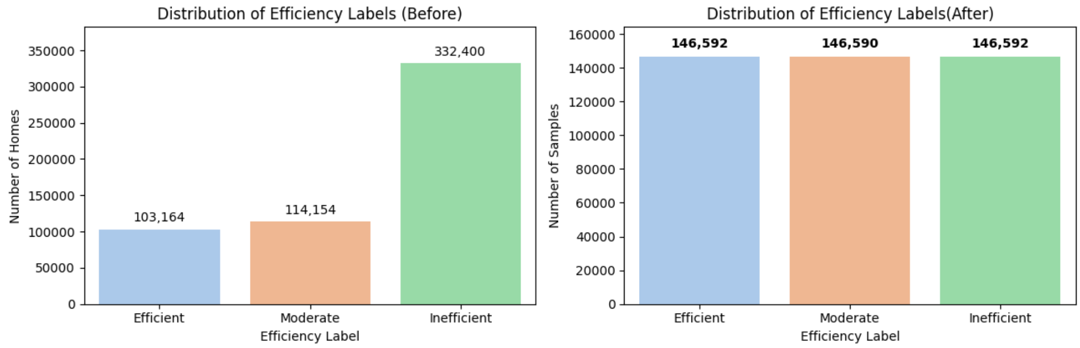

Figure 2: Efficiency label distribution before and after balancing.

In figure 2 the "before" plot highlights the severe skew toward *Inefficient* homes, while the "after" plot confirms a balanced dataset. This ensures fairer classification and more robust learning across all classes.

A correlation heat map was drawn that identified the relationship between features and the target variable `energy_per_sqft`. The factors that were least negatively correlated with energy intensity include CO_2 emissions and total energy use and the area of floor that was least positively correlated, indicating that the bigger the home the more energy efficient it seems to be per square foot.

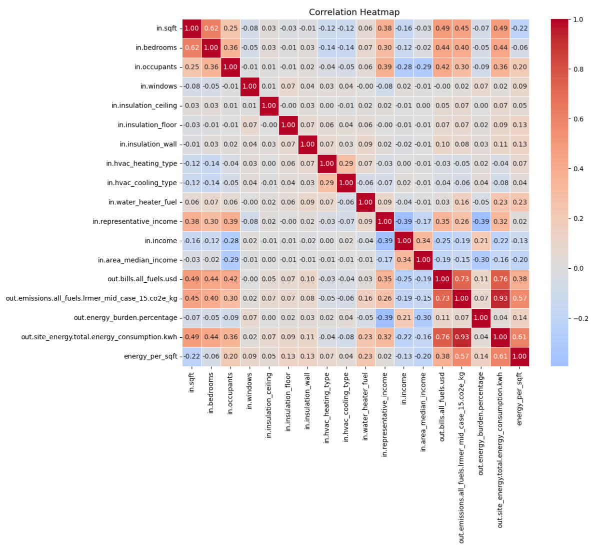

Figure 3: Correlation heatmap of features and target variable.

In figure 3 The positive relationships are overwhelmed with Emissions related variables; another strong negative relationship is building size with normalized energy use. All the selected predictors are retained because low or moderate multicollinearity is considered.

After the exploratory analysis,Three Classification models, namely Logistic Regression, Decision Tree Classifier and Random Forest Classifier were considered. Accuracy and macro-averaged F1-score were used to measure performance in order to achieve balanced performance across classes. Random Forest Classifier made the best performance (Accuracy = 87%, F1-score = 0.88), followed by decision tree (Accuracy = 87%, F1 = 0.87) and then Logistic Regression (Accuracy = 81%, F1 = 0.82). Mehmood et al. (2023) managed to report a slightly lower percentage of accuracy (85%) after employing gradient boosting.

Table 2: Classification model performance on the test set.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.81	0.82	0.81	0.82
Decision Tree	0.87	0.87	0.87	0.87
Random Forest	0.87	0.88	0.87	0.88

In Table 4 we can clearly see that the Random Forest model continues to outperform the easier baseline models (i.e., Logistic Regression, and Decision Tree) which implies the power of ensemble methods in such multi-class classification problems. Random Forest is a method of combining predictions of several decision trees, therefore it lowers the variance and alleviates overfitting and more complex decision boundaries. This leads to an increased overall accuracy, precision, recall and F1-scores, especially in cases where a linear separability of classes is not present.

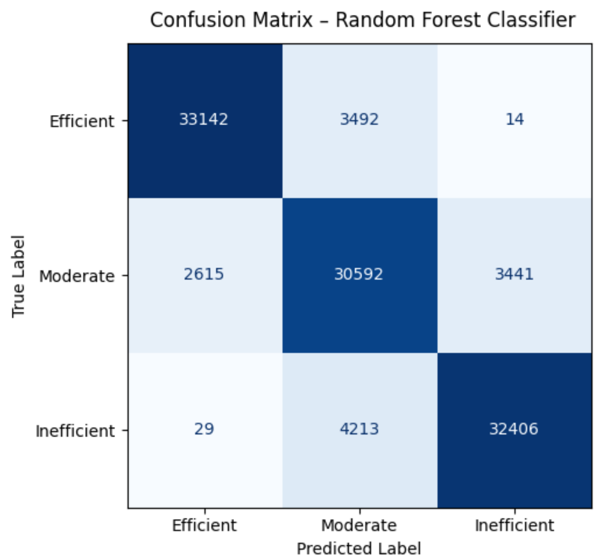

Figure 4: Confusion matrix for Random Forest Classifier.

A balanced accuracy in all three categories of efficiency based efficiency categories in that the model equally demonstrates to be effective in predicting an efficient, moderate and an inefficient home residence with the same result. The situation can be considered as relatively low when it comes to misclassifications, and there is no major ambiguity between the Efficient and the Inefficient classes which are classes that represent both extremities in terms of energy performance. The greatest percentage of errors are in the Moderate category, with similarity of feature characteristics with Efficient and Inefficient houses presenting a greater difficulty in clearly classifying the house. This is a balanced performance indicative of how the model generalizes quite well and does not favor any particular class.

For the regression task, Linear Regression, Decision Tree Regressor, and Random Forest Regressor were tested on predicting `energy_per_sqft`. Random Forest achieved $R^2 = 0.9426$, MAE = 1.4937 kWh/sqft, and MSE = 6.0318 kWh/sqft^2. Decision Tree achieved $R^2 = 0.8885$, while Linear Regression lagged with $R^2 = 0.6902$. These results outperform the 0.91 R^2 in Powroźnik & Szcześniak Powroźnik and Szcześniak (2024) and 0.78 in Fumo and Biswas (2015).

Table 3: Regression model performance on the test set.

Model	R^2	MAE	MSE
Linear Regression	0.6902	4.0270	32.5272
Decision Tree	0.8885	1.9828	11.7127
Random Forest	0.9426	1.4937	6.0318

Random Forest model has the most details of predication accuracy, which is symbolized by the largest R^2 , and the lowest error rates, as both MAE and MSE is far better than the other models. These findings show the resilience of the ensemble learning technique, and multiple decision trees can mitigate overfitting, improves the modeling of non-linear complex relationships and displays more stable and reliable predictions than using only a single model like Linear Regression or Decision Tree.

Figure 5: Comparison of regression model performance.

Figure 5 emphasizes the visual representation of Random Forest's clear advantage in predictive accuracy.

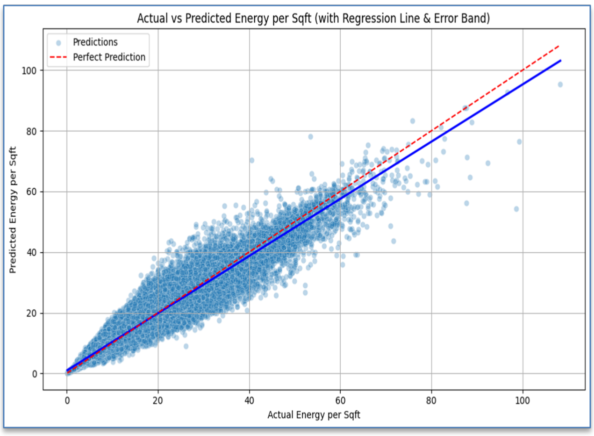

Figure 6: Actual vs predicted plot for Random Forest Regressor.

The plot demonstrates a strong alignment between predicted and actual values, with minimal dispersion from the diagonal, indicating high predictive reliability

To ensure interpretability, SHAP (SHapley Additive Explanations) was used to quantify the impact of each feature on model predictions. In the regression context, CO_2 emissions and floor area were the most influential features, followed by heating and cooling system types, insulation levels, and household income metrics. This ranking aligns with findings from Cui et al. (2024), where emissions, building size, and thermal envelope characteristics were also identified as dominant factors. Such agreement strengthens confidence in the model's validity and supports its potential use in targeted retrofit planning.

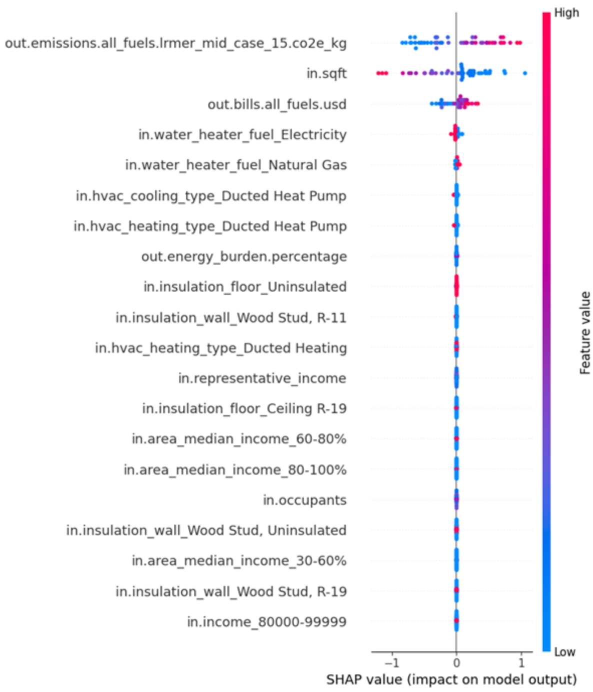

Figure 7: SHAP force plot for an individual prediction.

Based on the SHAP summary plot, the carbon emissions and the home size are the two factors that have the biggest impact on the energy efficiency classifications affecting the model predictions strongly and negatively. The energy bills or the type of heating or cooling systems including electric water heaters, ducted heat pumps, also contribute a great deal with higher values corresponding to the less efficient classifications. The characteristics that lead to the predictions are the quality of insulation, specifically uninsulated floors, and the socioeconomic indicators such as the representative income and area median income that have relatively lesser influence. All in all, the findings point out that a discussion on physical building characteristics along with household energy/emissions profiles are key elements in the decision-making process of the model.

7 Discussion,Conclusion and Future Work

7.1 Discussion

The results of the study show that the ensemble-based machine learning algorithms, specifically, Random Forest Classifier and Regressor, will give excellent results on making predictions concerning residential energy efficiency. The R^2 value on conducted regression task was 0.94, which is higher than what was provided by Powroźnik and Szcześniak (2024), and Fumo and Biswas (2015) (0.78). Random Forest achieved 87% accuracy in the classification problem together with a macro-averaged F1-score of 0.88, which is just a bit higher than Mehmood et al. (2023) at 85%. The balance of the dataset could be considered an advantage since the challenge of bias in the classes was discovered in Mahmood and Asif (2024), and SHAP values provided better understanding because they precisely establish the factors shaping the predictions.

Industry-leading strengths lie in using a diverse set of features including structure, insulation, mechanical, and socioeconomic variables and to incorporate the AI of explainability in order to propose a specific retrofit recommendation. There are however limitations. Without cross-validation and tuning of hyperparameters, potential performance improvement may be unrealised. The ResStock dataset is rather large, but simulation-derived and does not cover all the life-like variances. The retrofit simulation module has not been finalised and so the capability to test improvement scenarios against testing is not possible directly.

The results would be most relevant to U.S. style residential construction of roughly equivalent climate and construction trends as in the ResStock data. It will need local data re-training in the other areas and additionally using higher resolution inputs like occupancy and appliance usage might enhance model accuracy.

7.2 Conclusion & Future Work

The purpose of this study was to improve the accuracy of the residential energy efficiency prediction considering the accuracy and interpretations of machine learning models. The goals were to (1) form classification and regression models, (2) use a wide range of features rather than enforce more powerful predictive precision, and (3) incorporate explainability methods to locate critical driving forces of predictions. All the goals were met. The Random Forest Classifier reached accuracy 87% and the macro F1- score 0.88, whereas the Regressor obtained $R^2 = 0.94$, low values of MAE and MSE. CO_2 emissions, floor area, and the type of heating/cooling system were found to be the most significant predictors in the SHAP analysis consistent with the other charges.

These findings indicate that ensembles of balanced datasets and heterogeneous feature sets would be referenced to create accurate and transparent models. Decision-making by policymakers, utilities, and homeowners with respect to the optimal retrofit measure is achievable with the aid of such models.Overall, this research contributes to advancing the United Nations Sustainable Development Goal 7 (Affordable and Clean Energy) by providing a data-driven, interpretable framework for improving residential energy efficiency, thereby supporting the transition toward cleaner energy use and reduced environmental impact.

Future research will focus on:

1. Finishing the retrofit simulation module and see the possible energy reduction through smart upgrading like insulation and replacement of systems.

2. Model optimisation in two steps: use cross-validation and hyperparameter tuning to optimise performance and generalisability further.

3. Increase in size and localisation of the datasets - allowing the approach to be used in other climates and housing stocks using a variety of datasets available (e.g. real-world sensor, occupancy and appliance usage data).

In the long-term, this might see such models incorporated into smart home management systems where the continuous monitoring of energy consumption, along with subsequent dynamic retrofit suggestions would be able to create a closed feedback loop between quantifying performance and making energy-saving measures accordingly

References

Aghili, S. A., Rezaei, A. H. M., Tafazzoli, M., Khanzadi, M. and Rahbar, M. (2025), 'Artificial intelligence approaches to energy management in hvac systems: A systematic review', *Buildings* **15**, 1008.

Ali, U., Bano, S., Shamsi, M. H., Sood, D., Hoare, C., Zuo, W., Hewitt, N. and O'Donnell, J. (2024), 'Urban building energy performance prediction and retrofit analysis using data-driven machine learning approach', *Energy and Buildings* **303**, 113768.

Alzoubi, Y. I. and Mishra, A. (2024), 'Green artificial intelligence initiatives: Potentials and challenges', *Journal of Cleaner Production* **468**, 143090.

Asamoah, P. and Shittu, E. (2025), 'Evaluating the performance of machine learning models for energy load prediction in residential hvac systems', *Energy and Buildings* **334**, 115517.
URL: *https://www.sciencedirect.com/science/article/pii/S0378778825002476*

Balti, N., Vrigneau, B. and Scalart, P. (2024), 'Occupancy detection for enhanced energy disaggregation', *Procedia Computer Science* **246**, 529–537.

Bozorgi, M., Tasnim, S. H. and Mahmud, S. (2025), 'Machine learning-driven hybrid cooling system for enhanced energy efficiency in multi-unit residential buildings', *Energy and Buildings* **336**, 115613.

Cui, X., Lee, M., Koo, C. and Hong, T. (2024), 'Energy consumption prediction and household feature analysis for different residential building types using machine learning and shap', *Energy and Buildings* **309**, 113997.
URL: *https://www.sciencedirect.com/science/article/pii/S0378778824001130*

Dell'Anna, F. (2025), 'Machine learning framework for evaluating energy performance certificate (epc) effectiveness in real estate: A case study of turin's private residential market', *Energy Policy* **198**, 114407.

Fumo, N. and Biswas, M. R. (2015), 'Regression analysis for prediction of residential energy consumption', *Renewable and Sustainable Energy Reviews* **47**, 332–343. [Online; accessed 7-Aug-2025].

Gorzałczany, W. and Rudziński, F. (2024), 'Energy consumption prediction in residential buildings—an accurate and interpretable machine learning approach combining fuzzy systems with evolutionary optimization', *Energies* **17**(13), 3242.
URL: *https://www.mdpi.com/1996-1073/17/13/3242*

Husseini, F. E., Noura, H. N., Salman, O. and Chahine, K. (2025), 'Machine learning in smart buildings: A review of methods, challenges, and future trends', *Applied Sciences* **15**(14), 7682.

International Energy Agency (2022), 'Heating'. [Online; accessed 7-Aug-2025].
URL: *https://www.iea.org/reports/heating*

Khan, Z., Hussain, T., Ullah, A., Rho, S., Lee, M. and Baik, S. (2020), 'Towards efficient electricity forecasting in residential and commercial buildings: A novel hybrid cnn with a lstm-ae based framework', *Sensors* **20**(5), 1399.

Lee, S. et al. (2022), 'Power management in smart residential building with deep learning model for occupancy detection by usage pattern of electric appliances', *Energy Reports* **8**, 3365–3375.
URL: *https://arxiv.org/pdf/2209.11520*

Liu, J. and Chen, J. (2025), 'Applications and trends of machine learning in building energy optimization: A bibliometric analysis', *Buildings* **15**(7), 994.
URL: *https://www.mdpi.com/2075-5309/15/7/994*

Lotfipoor, A., Patidar, S. and Jenkins, D. P. (2024), 'Deep neural network with empirical mode decomposition and bayesian optimisation for residential load forecasting', *Expert Systems with Applications* **237**, 121355.

Mahmood, T. and Asif, M. (2024), 'Prediction of energy efficiency for residential buildings using supervised machine learning algorithms', *Energies* **17**(19), 4965.

Mehmood, A., Lee, K.-T. and Kim, D.-H. (2023), 'Energy prediction and optimization for smart homes with weather metric-weight coefficients', *Sensors* **23**(7), 3640.
URL: *https://www.mdpi.com/1424-8220/23/7/3640*

Moosbrugger, L., Seiler, V., Wohlgenannt, P., Hegenbart, S., Ristov, S., Eder, E. and Kepplinger, P. (2025), 'Load forecasting for households and energy communities: Are deep learning models worth the effort?'.
URL: *https://arxiv.org/abs/2501.05000*

Nandkeolyar, S., Ray, P., Puhan, P., Panda, G. and Panda, R. (2025), 'A machine learning-based hybrid deep neural network approach for adaptive demand side management using community-level battery storage systems', *IEEE Transactions on Industry*

Applications pp. 1–10.
URL: *https://ieeexplore.ieee.org/document/10989556*

Nigar, N. et al. (2024), 'Improving demand-side energy management with energy advisor using machine learning', *Journal of Electrical and Computer Engineering* .
URL: *https://www.researchgate.net/publication/385884435*

Peplinski, M. et al. (2024), 'A machine learning framework to estimate residential electricity demand based on smart meter electricity, climate, building characteristics, and socioeconomic datasets', *Applied Energy* **357**, 122413.
URL: *https://www.sciencedirect.com/science/article/pii/S0306261923017774*

Powroźnik, P. and Szcześniak, P. (2024), 'Predictive analytics for energy efficiency: Leveraging machine learning to optimize household energy consumption', *Energies* **17**(2), 459.
URL: *https://www.mdpi.com/1996-1073/17/23/5866*

Qiao, Q. and Yunusa-Kaltungo, A. (2023), 'A hybrid agent-based machine learning method for human-centred energy consumption prediction', *Energy and Buildings* **283**, 112797.

Raza, A., Jingzhao, L., Ghadi, Y., Adnan, M. and Ali, M. (2024), 'Smart home energy management systems: Research challenges and survey', *Alexandria Engineering Journal* **92**, 117–170.

Rehman, U., Faria, P., Gomes, L. and Vale, Z. (2025), 'Future of energy management models in smart homes: A systematic literature review', *Process Integration and Optimization for Sustainability* .
URL: *https://link.springer.com/article/10.1007/s41660-025-00506-x*

Rempi, A. et al. (2025), 'Explainable retrofit prediction for residential buildings under data scarcity'.
URL: *https://arxiv.org/abs/2504.06055*

Roach, C. (2020), 'Estimating electricity impact profiles for building characteristics using smart meter data and mixed models', *Energy and Buildings* **211**, 109686.

Sherif, T., Katia, R., Nguyen, M., Ma, N. and Rakha, T. (2025), 'Localizing urban building energy modeling (ubem) through inclusive microclimate and socioeconomic data', *Applied Energy* **383**, 125342.

Steephen, S., R, S. and N, N. (2022), A review on non-intrusive load monitoring using deep learning, *in* '2022 International Conference on Innovations in Science and Technology for Sustainable Development (ICISTSD)', pp. 381–385.
URL: *https://ieeexplore.ieee.org/document/10010467*

Thorve, S., Baek, Y., Swarup, S., Mortveit, H., Marathe, A., Vullikanti, A. and Marathe, M. (2023), 'High resolution synthetic residential energy use profiles for the united states', *Scientific Data* **10**.
URL: *https://www.nature.com/articles/s41597-022-01914-1*

Uddin, M. et al. (2024), 'Predicting occupant energy consumption in different indoor layout configurations using a hybrid agent-based modeling and machine learning approach', *Energy and Buildings* **328**, 115102.
URL: *https://www.sciencedirect.com/science/article/pii/S0378778824012180*

Yesilyurt, H., Dokuz, Y. and Dokuz, A. (2024), 'Data-driven energy consumption prediction of a university office building using machine learning algorithms', *Energy* **310**, 133242.
URL: *https://www.sciencedirect.com/science/article/pii/S0360544224030184*

Zoubir, Z., ErRetby, H., EsSakali, N., Souldi, A. and Mghazli, M. O. (2024), 'Towards sustainable buildings: Predictive modeling of energy consumption with machine learning', *Procedia Computer Science* **236**, 59–66.
URL: *https://www.sciencedirect.com/science/article/pii/S1877050924010202*