# Model Failure or Data Corruption? Exploring Inconsistencies in Building Energy Ratings with Self-Supervised Contrastive Learning

**Qian Xiao, Dan Liu & Kevin Credit**
ADAPT Centre
O'Reilly Institute, Trinity College 25 Westland Row, Dublin, D02 PN40
{qian.xiao, dan.liu, kevin.credit}@adaptcentre.ie

## Abstract

Building Energy Rating (BER) stands as a pivotal metric, enabling building owners, policymakers, and urban planners to understand the energy-saving potential through improving building energy efficiency. As such, enhancing buildings' BER levels is expected to directly contribute to the reduction of carbon emissions and promote climate improvement. Nonetheless, the BER assessment process is vulnerable to missing and inaccurate measurements. In this study, we introduce CLEAR, a data-driven approach designed to scrutinize the inconsistencies in BER assessments through self-supervised contrastive learning. We validated the effectiveness of CLEAR using a dataset representing Irish building stocks. Our experiments uncovered evidence of inconsistent BER assessments, highlighting measurement data corruption within this real-world dataset.

## 1 Introduction

The building sector is responsible for nearly 40% of energy-related CO2 emissions (Dahlström et al., 2022). Improving building energy efficiency, especially for existing ones, is key to combating climate change for many countries (Egan et al., 2023). Building energy efficiency assessment plays a pivotal role in guiding decisions for both retrofitting existing buildings and designing new ones (Coyne & Denny, 2021; Energy & Engineers). The assessment results stand as cornerstone metrics in shaping various policies for guiding the public to address energy poverty and steer climate action. For example, in Europe, buildings' energy efficiency is factored into mortgage interests rates for the homeowners (Billio et al., 2022). New dwellings in many EU countries are also required to comply with strict energy efficiency regulations (Energy & Engineers). Despite its significance, the assessment process of building energy efficiency in which professionals involved is susceptible to missing values and faulty measurements as the result of negligent errors (Ali et al., 2020). Consequently, such processes are prone to partial data corruption, leading to potential substantial inaccuracies in the assigned energy ratings.

Ensuring the reliability and transparency of the Building Energy Rating (BER) process is crucial for upholding the integrity of energy efficiency assessments. To tackle this challenge, previous research on Ireland's BER data has emphasized the potential benefits of integrating data-driven approaches (Ali et al., 2020). While such approach is promising, the machine learning models examined in (Ali et al., 2020) exhibited a significant performance drop when predicting fine-grained rating level categories. Table 1 presents such performance drops observed in both the Random Forest model and the MLP (multi-layer perceptron) deep learning model. Various factors could potentially contribute to these performance drops, including models' limited generalizability, sparsity of data distribution, or poor data quality. Currently, there is still a lack of understanding of the actual reasons behind such observations, which, in turn, hamper the potential of data-driven approaches to facilitate the BER assessment process.

This study aims to investigate such observations using self-supervised deep learning approaches. Specifically, we propose CLEAR, a self-supervised learning approach that employs Contrastive

| BER Levels | Performance Measure | Random Forest | MLP |
|---|---|---|---|
| A1, A2, ..., E1, E2, F, G | Accuracy (%) | 62.8 | **69.5** |
| | Macro F1 (%) | 63.1 | **63.9** |
| A, B, C, D, EFG | Accuracy (%) | 76.1 | **88.6** |
| | Macro F1 (%) | 75.8 | **88.9** |

Table 1: Performance Comparison of Models for BER Prediction with Different Level Granularity



(a) Five BER Levels  (b) Fifteen BER Levels

Figure 1: 2D PCA Representations in the Latent Spaces of Scarf Models

Learning for Energy Assessment Rating evaluation, aiming to provide explainability and improve the reliability of assessment data. Our approach's training process relies entirely on building measurement data, eliminating the need for assessment labels in training. This approach is preferable for our investigation, because the modelling process doesn't require assessment labels that are subjective to individual assessors' judgment. Moreover, our approach utilizes the SCARF model (Bahri et al., 2021) for contrastive learning, providing an advantage in mitigating the potential presence of data noise in modeling. In the modeling process, potential existing data corruption is leveraged to facilitate the randomized feature corruption process. This, in turn, may even enhance the model's generalizability and result in better representations of buildings.

With a Irish building stocks dataset, we discovered evidence of clear inconsistencies in energy ratings. The results demonstrate that buildings with similar feature values were given very different rating levels for their BER assessments in this dataset. By revealing these inconsistencies in building energy assessments, we identified significant faulty values in the building features. Consequently, such data corruption may largely limit the potential of data-driven approaches for BER assessment. This explains the observation that most of these approaches failed to achieve satisfactory predictions when the rating was at a fine-grained level.

## 2 APPROACH

Our approach CLEAR consists of two steps. First, we adopt self-supervised contrastive learning via SCARF modelling (Bahri et al., 2021) to extract representations for buildings. For contrastive learning, we generate positive pairs by randomly selecting and corrupting a subset of features. The negative pairs are generated by contrasting them with other records in a batch. After training, we extract the latent representation of each building with SCARF's encoder. Second, we explore the model's latent space to examine the rating inconsistency based on latent representations. Specifically, we visualize latent representations via PCA compression in 2D or 3D space. Close proximity of building representations in the latent space indicates similarities in their measured feature values. We can visually identify rating inconsistencies by observing close representations with different BER rating levels in the compressed PCA space. Once the reference buildings are located, we proceed to calculate the nearest neighboring buildings in the derived high-dimensional latent space.
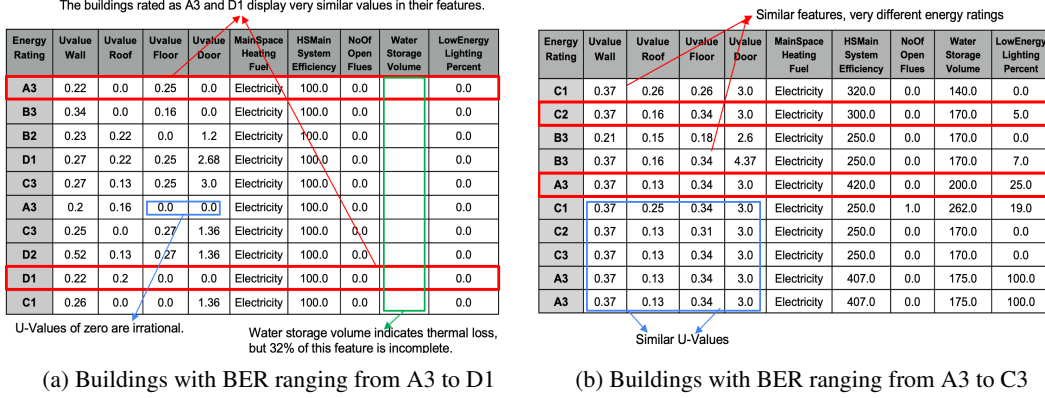
## 3 EXPERIMENTS

The buildings rated as A3 and D1 display very similar values in their features.

| Energy Rating | Uvalue Wall | Uvalue Roof | Uvalue Floor | Uvalue Door | MainSpace Heating Fuel | HSMain System Efficiency | NoOf Open Flues | Water Storage Volume | LowEnergy Lighting Percent |
|---|---|---|---|---|---|---|---|---|---|
| A3 | 0.22 | 0.0 | 0.25 | 0.0 | Electricity | 100.0 | 0.0 | | 0.0 |
| B3 | 0.34 | 0.0 | 0.16 | 0.0 | Electricity | 100.0 | 0.0 | | 0.0 |
| B2 | 0.23 | 0.22 | 0.0 | 1.2 | Electricity | 100.0 | 0.0 | | 0.0 |
| D1 | 0.27 | 0.22 | 0.25 | 2.68 | Electricity | 100.0 | 0.0 | | 0.0 |
| C3 | 0.27 | 0.13 | 0.25 | 3.0 | Electricity | 100.0 | 0.0 | | 0.0 |
| A3 | 0.2 | 0.16 | 0.0 | 0.0 | Electricity | 100.0 | 0.0 | | 0.0 |
| C3 | 0.25 | 0.0 | 0.27 | 1.36 | Electricity | 100.0 | 0.0 | | 0.0 |
| D2 | 0.52 | 0.13 | 0.27 | 1.36 | Electricity | 100.0 | 0.0 | | 0.0 |
| D1 | 0.22 | 0.2 | 0.0 | 0.0 | Electricity | 100.0 | 0.0 | | 0.0 |
| C1 | 0.26 | 0.0 | 0.0 | 1.36 | Electricity | 100.0 | 0.0 | | 0.0 |

U-Values of zero are irrational.

Water storage volume indicates thermal loss, but 32% of this feature is incomplete.

Similar features, very different energy ratings

| Energy Rating | Uvalue Wall | Uvalue Roof | Uvalue Floor | Uvalue Door | MainSpace Heating Fuel | HSMain System Efficiency | NoOf Open Flues | Water Storage Volume | LowEnergy Lighting Percent |
|---|---|---|---|---|---|---|---|---|---|
| C1 | 0.37 | 0.26 | 0.26 | 3.0 | Electricity | 320.0 | 0.0 | 140.0 | 0.0 |
| C2 | 0.37 | 0.16 | 0.34 | 3.0 | Electricity | 300.0 | 0.0 | 170.0 | 5.0 |
| B3 | 0.21 | 0.15 | 0.18 | 2.6 | Electricity | 250.0 | 0.0 | 170.0 | 0.0 |
| B3 | 0.37 | 0.16 | 0.34 | 4.37 | Electricity | 250.0 | 0.0 | 170.0 | 7.0 |
| A3 | 0.37 | 0.13 | 0.34 | 3.0 | Electricity | 420.0 | 0.0 | 200.0 | 25.0 |
| C1 | 0.37 | 0.25 | 0.34 | 3.0 | Electricity | 250.0 | 1.0 | 262.0 | 19.0 |
| C2 | 0.37 | 0.13 | 0.31 | 3.0 | Electricity | 250.0 | 0.0 | 170.0 | 0.0 |
| C3 | 0.37 | 0.13 | 0.34 | 3.0 | Electricity | 250.0 | 0.0 | 170.0 | 0.0 |
| A3 | 0.37 | 0.13 | 0.34 | 3.0 | Electricity | 407.0 | 0.0 | 175.0 | 100.0 |
| A3 | 0.37 | 0.13 | 0.34 | 3.0 | Electricity | 407.0 | 0.0 | 175.0 | 100.0 |

Similar U-Values

(a) Buildings with BER ranging from A3 to D1          (b) Buildings with BER ranging from A3 to C3

Figure 2: Inconsistent Ratings & Measurement Data Corruption

**Identifying Rating Inconsistency in the Latent Space.** For each building in the dataset[1], we extract the latent representation in a 32-dimensional vector. Subsequently, we derive a 2D visualization by compressing the latent representations further to two dimensions using PCA. Figure 1 displays the 2D visualization of latent spaces for different BER granularities. As shown in Figure 1a, buildings with five coarse BER levels are spread across the latent space in sequence, aligned with the order of BER levels. Those with the same BER level are predominantly clustered together. However, in Figure 1b, the distribution of latent representations reveals that buildings with adjacent BER levels mix together, and the boundaries of fine-grained rating levels for each coarse level (e.g., 'B1', 'B2', 'B3' for coarse rating 'B') are not clear. This indicates inconsistent ratings for buildings with similar feature values, especially evident in neighboring levels such as the group ('B1', 'B2', 'B3'), the group ('C1', 'C2', 'C3'), and the group ('D1', 'D2'). This observation aligns with the findings in the confusion matrix shown in Figure 4 in the appendix for the MLP model.

**Data Corruption in the Inconsistent BER Rating Records.** The positioning of buildings' representations in the latent space reveals the distribution of buildings with different features. Proximity in the latent space signifies similarity in features between two buildings. Consequently, identifying similar buildings with different BER levels becomes evident in the latent space. For instance, Figure 2a showcases ten buildings resembling the reference building with BER level 'A3' (the first row in the table). These buildings are derived by calculating the top ten closest representations in the 32-dimensional latent space. It is notable that these buildings share very similar feature values for U-values of the wall, roof, door, window, and floor—features among those of top importance identified by the decision tree algorithm. Additionally, their heating systems possess identical specifications. However, their ratings range from 'A3' to 'D1' in the assessment. In addition, we also found out substantial abnormal values in lighting and water storage highlight potential data corruption in their measurements. As another example, Figure 2a presents another group of similar buildings, with a specific focus on two buildings (2nd row and 5th row) that exhibit almost identical feature values across all dimensions but are assessed differently as 'C2' and 'A3', respectively. The five buildings at the bottom also display similar U-values in all aspects. For reference purpose, details of value ranges for U-values can be presented in the box-plots (see Figure 3 in the appendix).

## 4   CONCLUSION

In this study, we introduce a data-driven approach based on self-supervised contrastive learning to identify inconsistencies in BER assessments. Using an Irish building stock dataset, our experiments show the effectiveness of the proposed approach in detecting rating inconsistency as well as measurement data corruption in the BER assessment process.

---

[1]The description of the dataset we use and our experiments configuration can be found in the appendix.

REFERENCES

Usman Ali, Mohammad Haris Shamsi, Mark Bohacek, Karl Purcell, Cathal Hoare, Eleni Mangina, and James O'Donnell. A data-driven approach for multi-scale gis-based building energy modeling for analysis, planning and support decision making. *Applied Energy*, 279:115834, 2020.

Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. Scarf: Self-supervised contrastive learning using random feature corruption. In *International Conference on Learning Representations*, 2021.

Monica Billio, Michele Costola, Loriana Pelizzon, and Max Riedel. Buildings' energy efficiency and the probability of mortgage default: The dutch case. *The Journal of Real Estate Finance and Economics*, 65(3):419–450, 2022.

Bryan Coyne and Eleanor Denny. Mind the energy performance gap: testing the accuracy of building energy performance certificates in ireland. *Energy efficiency*, 14(6):57, 2021.

Lukas Dahlström, Tor Broström, and Joakim Widén. Advancing urban building energy modelling through new model components and applications: A review. *Energy and Buildings*, 266, 2022.

Hilary Egan, Clement Fouquet, and Chioke Harris. Machine learning for advanced building construction. In *ICLR 2023 Workshop on Tackling Climate Change with Machine Learning*, 2023.

Emerald Energy and Esbensen Consulting Engineers. Energy efficiency regulations for new dwellings and options for improvement.

## A  APPENDIX

### A.1  DATA PREPROCESSING & FEATURE EXTRACTION

The data preprocessing consists of three steps. The first step is data cleaning. In this step, we use the interquartile range (IQR) technique to identify and address outliers. We then group the data by building type. For numerical features, missing values are imputed with the mean value specific to each building type. For categorical features, we utilized the most frequent value in each category for fill-ins. In the second step, we selected 40 features from the total 211 features in the original dataset. The selection is based on the feature importance analysis with decision tree algorithm, which identified the features with the most significant impact on energy ratings. During this process, we excluded compound features that can be derived by the actual energy ratings, such as total primary energy use for the dwelling and the dwelling's carbon dioxide emissions (known as $CO_2$ emissions). These 40 features are categorised into distinct groups, each describing different aspects of a building. These categories include building envelope features (e.g., the area of the wall, roof, and door), building fabric features (e.g., the U-values for the wall, roof, and door), characteristics of the heating system (e.g., the main heating system efficiency), hot water-related attributes (e.g., water storage volume), and spatial features (e.g., the county code). Lastly, in the third step, we applied standard scalers to standardize the numerical features and one-hot encoders to transform categorical features.

### A.2  DATASET

We use the Energy Performance Certificates (EPC) dataset [2] in this study, collected by the Sustainable Energy Authority of Ireland. This dataset contains 112,528 building assessment records. There are 15 different BER levels in total, ranging from A1 as the highest level to G as the lowest level.
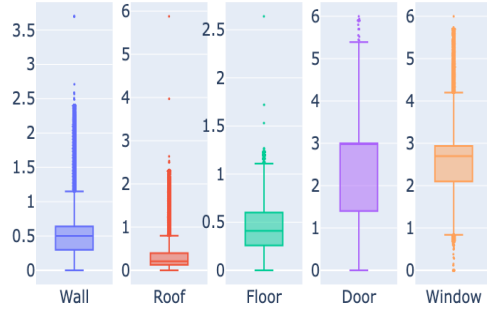
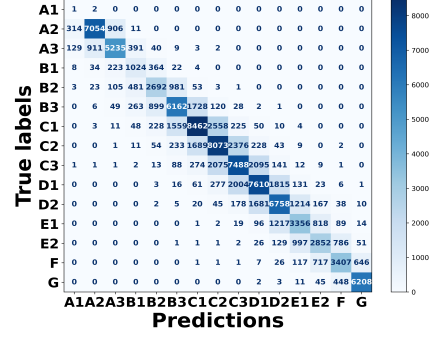Figure 3: Box-plots of U-Values



Figure 4: Confusion Matrix for MLP Predictions

## A.3 EXPERIMENTS CONFIGURATION

In our experiments, the SCARF model was trained for 15 epochs, with a batch size of 16 and a learning rate of 0.001. The model used in the experiments consists of three MLP blocks, each with a linear layer followed by ReLU activation. We set the dimension of the model's encoder to be 32. The proportion of randomly selected feature corruption is set to be 30%. We split the dataset into train, validation, and test sets by the ratio 80%, 10%, and 10%.

---

[2]https://ndber.seai.ie/BERResearchTool/ber/search.aspx