

Leveraging Self-Supervised Transformers for Heat Pump Consumption Classification Using the HEAPO Dataset

Hoang Thi Thu¹, Phi Dinh Huynh², Ta Trung Kien², Le Mai Nam², Le Hong Duy³, Pham Hong Thinh⁴,
 Nguyen Viet Hung^{2,*}

¹Posts and Telecommunications Institute of Technology, Hanoi, Vietnam

²East Asia University of Technology, Hanoi, Vietnam

³Faculty of Information Technology, CMC University, Hanoi, Vietnam

⁴Faculty of Engineering and Technology, Quy Nhon University, Quynhon, Vietnam

Article Info

Article history:

Received Dec 28, 2025

Revised Aug 22, 2026

Accepted Sep 9, 2026

Keywords:

Heat Pump Monitoring
 Smart-Meter Data
 Self-Supervised Pretraining
 Transformer Encoder
 Energy Consumption
 Classification
 HEAPO Dataset.

ABSTRACT

Heat pumps play a crucial role in reducing carbon emissions in residential heating; however, their electricity consumption varies significantly across households, making large-scale performance assessment challenging. Existing data-driven approaches often rely on fully supervised learning and require extensive labeled data, which are costly to obtain through on-site inspections. This paper investigates whether routinely collected smart meter and weather data can be leveraged to automatically classify household heat pump electricity consumption into three categories: low, normal, and rather high. We propose a self-monitoring transformer framework that combines self-supervised pretraining and supervised fine-tuning for time series classification. A transformer encoder is first pre-trained on multi-day electricity consumption windows from largely unlabeled households using a reconstruction-based objective. The learned representations are then fine-tuned with a lightweight fully connected classifier on a limited labeled subset of the HEAPO dataset. Experimental results demonstrate that the proposed approach consistently outperforms state-of-the-art baselines, including RNN-, CNN-, and transformer-based time series models, achieving approximately 93% classification accuracy, an F1 score above 0.9, and a recovery rate close to 91%. These results indicate that self-supervised representation learning on smart meter data provides an effective and scalable solution for monitoring heat pump electricity consumption and supporting inspection planning.

Copyright © 2026 Institute of Advanced Engineering and Science.
 All rights reserved.

Corresponding Author:

Nguyen Viet Hung

International Training and Cooperation Institute, East Asia University of Technology, Hanoi, Vietnam

Email: hungnv@eaut.edu.vn

1. INTRODUCTION

Heating and cooling remain one of the largest end uses of energy worldwide, accounting for around half of global final energy consumption and a sizable share of energy-related CO₂ emissions [1–3]. In residential buildings, heat pumps are increasingly deployed as a key technology for electrifying space heating and domestic hot water while enabling the use of low-carbon electricity. Yet many field deployments operate far below their theoretical efficiency because of sizing errors, poor configuration, or unnoticed faults, which translate into higher bills for households and additional stress on the power system [4]. Identifying such problematic installations typically requires on-site visits by trained consultants, but visiting every home is infeasible at scale. In practice, a central question emerges: *Which heat-pump households should be prioritized for inspection based*

on their electricity consumption patterns?

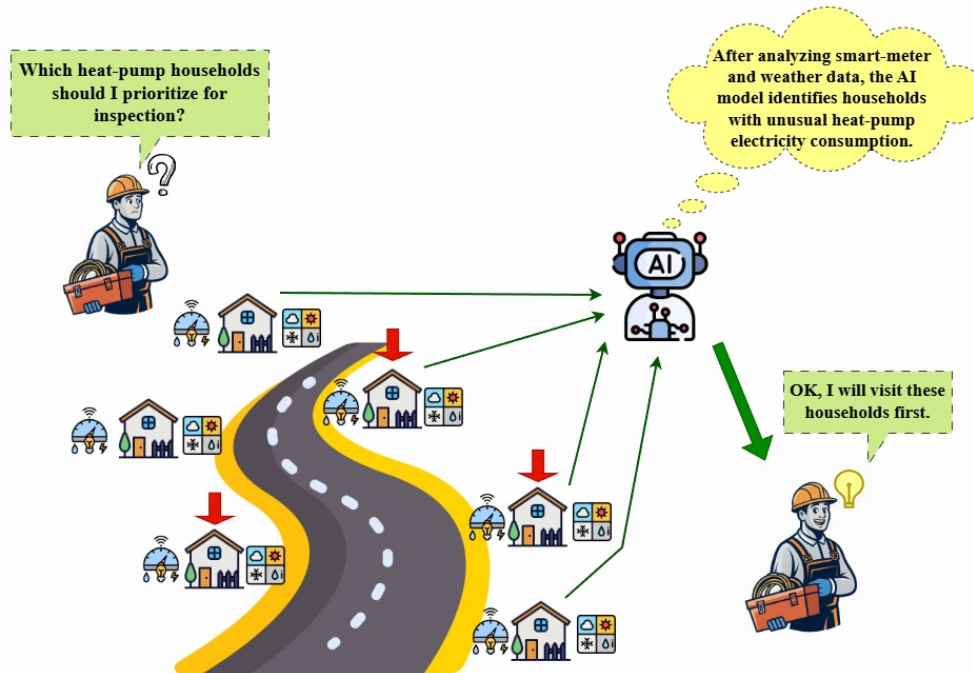


Figure 1. Illustrating the concept of AI-powered pre-selection of homes for heat pump inspections. Instead of manually guessing which homes to inspect, an energy consultant uses an AI model that analyzes smart meter and weather data to identify homes with unusually high heat pump electricity consumption and prioritizes inspections.

Fig. 1 illustrates the motivation of this study in modern energy-aware residential environments, where technicians must monitor large numbers of homes equipped with heat pumps and smart meters. Although multi-year electricity consumption and weather data are continuously collected, identifying inefficient or abnormal heat pump operation remains challenging. To address this problem, our AI model analyzes large-scale smart-meter time-series data in parallel and employs a Transformer-based architecture to learn robust consumption representations, which have proven effective in capturing long-range dependencies across domains [5]. In contrast, shallow learning approaches exhibit limited scalability and performance for complex classification tasks [6]. As shown in the figure, residences with unusually high heat pump consumption are highlighted as high-priority candidates for inspection, enabling experts to focus their efforts where interventions are most impactful. Building on recent advances in intelligent system monitoring [7–11], this work demonstrates how large, noisy data streams can be transformed into actionable decision support for energy-efficient operation.

The recently released HEAPO dataset makes this setting realistic. HEAPO provides smart-meter electricity data for 1 408 Swiss households with heat pumps, recorded at 15-minute and daily resolution between late 2018 and early 2024, together with household metadata, local weather from multiple stations, and 410 on-site inspection protocols produced by professional energy consultants [4, 12]. Each protocol includes a qualitative assessment of the heat pump's electricity consumption (for example "normal" or "rather high"), which can serve as ground truth for model evaluation. This combination of long-horizon consumption data and sparse expert labels turns HEAPO into a natural testbed for data-driven pre-selection of inspection targets.

Prior work on building energy analytics and smart-meter data has explored various anomaly-detection and fault-diagnosis schemes, ranging from statistical baselines to machine-learning and deep-learning methods [13, 14]. Most approaches either detect short-term anomalies in aggregate load, or focus on disaggregating usage to individual appliances. At the same time, self-supervised learning (SSL) for time series has emerged as a powerful paradigm for exploiting large unlabeled sensor corpora by pretraining models on proxy tasks such as masked reconstruction or contrastive prediction, then fine-tuning on small labeled datasets [15]. However, to the best of our knowledge, there has been no systematic study of using SSL on multi-year smart-meter traces

to *rank* heat-pump households for inspection using labels derived from professional on-site protocols.

In this paper, we investigate the task of classifying household-level heat pump consumption on the HEAPO dataset and design a simple self-monitoring Transformer process for this task. First, we pre-train a Transformer encoder on hundreds of thousands of overlapping 7-day windows extracted from a 15-minute smart meter data series, supplemented with daily outdoor temperature data. During pre-training, adjacent segments of each window are randomly masked over time, and a lightweight reconstruction processor is trained to recover missing values from the unmasked context. In the second stage, the pre-trained encoder is reused as a feature extractor and fine-tuned along with a small classification processor on a subset of households with test labels, learning to predict three consumption classes (low, normal, fairly high) for each window. Subsequently, household-level predictions were obtained by aggregating the outputs at the window level, and we evaluated the resulting classifier against several baseline models using standard multi-class indices.

Our contribution comprises four aspects:

- We formulate a household-level three-class heat-pump electricity consumption classification task on the HEAPO dataset, combining smart-meter, weather, and inspection information.
- We propose a self-supervised masked-reconstruction framework that pretrains a transformer encoder on multi-day smart-meter and temperature windows using only unlabeled households.
- We fine-tune the pretrained encoder for consumption category prediction with limited labeled households and show that it outperforms non-pretrained transformers and simple feature-based baselines.
- Finally, we benchmark the proposed approach against several strong sequence-model baselines for time-series data, including RNN, LSTM, GRU, TCN, TimesNet and TS2Vec, to provide an objective comparison of classification performance.

The remainder of this paper is organized as follows. Section 2. reviews related work on smart-meter analytics, self-supervised learning for time series, and heat-pump monitoring. Section 3. describes the proposed preprocessing pipeline and the transformer-based pretraining and fine-tuning framework. Section 4. presents the experimental setup and evaluation results on the HEAPO dataset. Finally, Section 5. concludes the paper and outlines directions for future work.

2. RELATED WORK

Several empirical studies have investigated the discrepancy between nominal and in-situ performance of residential heat pumps. Large-scale field campaigns in Switzerland and other European countries report substantial variability in seasonal performance factors, often attributing inefficiencies to planning errors, commissioning issues, and suboptimal control strategies [16, 17]. Analyses of inspection protocols further reveal frequent so-called “soft faults” that increase electricity consumption without triggering system failures [18, 19]. More recent work explores model-based and data-driven fault diagnosis methods for specific heat pump configurations, but these studies typically rely on laboratory setups or small case studies rather than large residential fleets [20, 21].

The HEAPO dataset [4] extends this line of research by combining long-term smart-meter measurements from 1,408 households with detailed inspection reports and weather data. While the original HEAPO study focuses on dataset construction and rule-based screening strategies, it does not develop a dedicated machine-learning framework for predicting household-level consumption categories from time-series data. Related work in industrial energy analytics has shown that scalable smart-meter data processing and learning-based approaches are essential for monitoring and decision support in large energy systems [22, 23]. In parallel to data-driven anomaly detection, domain-specific literature heavily emphasizes physical heat pump coefficient of performance (COP) degradation modeling and weather-normalization techniques, such as degree-day normalization, to assess system efficiency. However, relying purely on physical models typically requires detailed parameters regarding building thermal inertia and envelope insulation, which are rarely available in large-scale smart-meter deployments. This limitation highlights the need for robust data-driven models that can bypass explicit physical modeling while still accounting for weather variations.

A broader body of literature addresses anomaly detection and consumption analysis in buildings using smart-meter data. Recent surveys highlight a shift toward deep learning and hybrid approaches, while also

emphasizing challenges such as noisy signals, non-stationarity, and limited labeled data [24, 25]. At the algorithmic level, many methods operate on daily or hourly load profiles to identify abnormal periods or events, using unsupervised or semi-supervised techniques [26, 27]. However, these approaches typically focus on detecting isolated anomalies rather than predicting household-level consumption categories informed by expert inspection outcomes.

In parallel, the time-series learning community has proposed increasingly powerful sequence models and representation learning techniques. Architectures such as Temporal Convolutional Networks, TimesNet, and contrastive representation models like TS2Vec have demonstrated strong performance across forecasting, classification, and anomaly detection benchmarks [28–30]. These methods motivate the use of deep temporal encoders that can capture multi-scale patterns in long consumption sequences.

Self-supervised learning has recently emerged as a particularly effective paradigm for time-series analysis under limited supervision. By leveraging pretext tasks such as masked reconstruction including recent advances in patch-level temporal masking that prevent trivial linear interpolation in highly autocorrelated data or contrastive objectives, transformer-based encoders pretrained on large unlabeled corpora produce transferable representations for downstream tasks. While applying transformer-based SSL to general building energy forecasting and time-series anomaly detection is well established [15, 27, 31], utilizing it specifically for heat-pump inspection pre-selection using highly subjective, qualitative inspection protocols presents a unique domain challenge. Therefore, the primary contribution lies not in designing a fundamentally novel SSL architecture, but in bridging this specific research gap: demonstrating how a tailored self-supervised masked-reconstruction framework can be optimized to extract actionable, household-level classification insights from noisy smart-meter data, weather information, and sparse expert labels.

3. PROPOSED METHOD

In this section, we outline the proposed pipeline. Figure 2 shows the overall procedure on the HEAPO dataset. Starting from the raw files, we perform basic cleaning and integration to build three aligned data streams per household: 15-minute smart-meter electricity use, daily outdoor temperature from the corresponding weather station, and on-site inspection reports when available. The smart-meter and temperature series are then sliced into overlapping 7-day sliding windows, which serve as the common input to all subsequent stages.

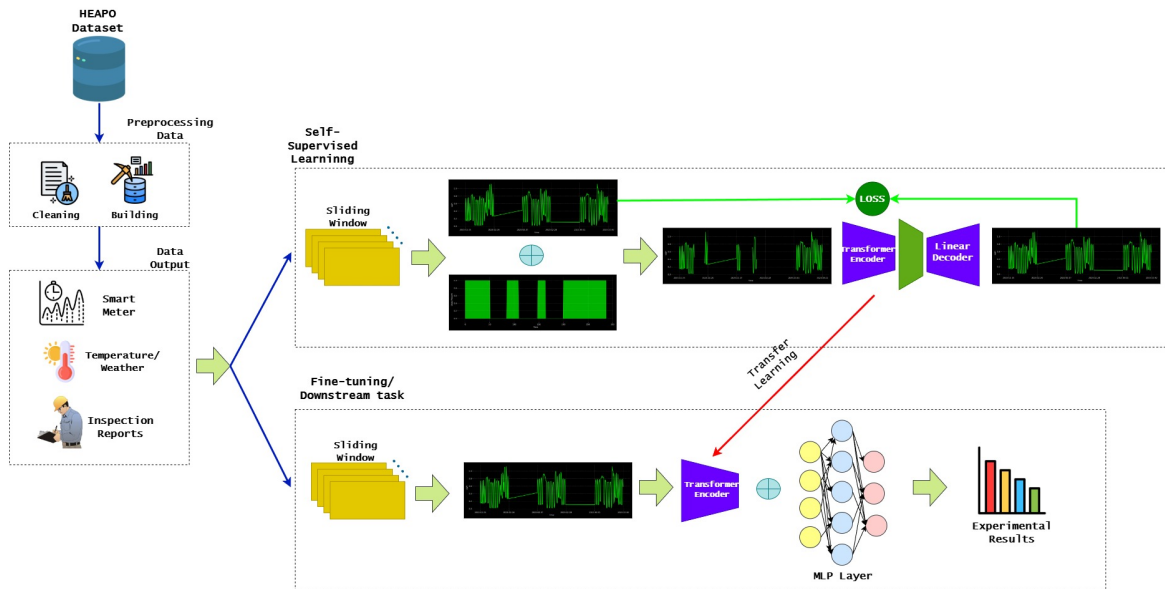


Figure 2. Overview of the HEAPO dataset construction process.

The upper branch of the figure corresponds to self-supervised pretraining. For each window we randomly mask a subset of time steps, feed the masked series into a transformer encoder, and attach a lightweight linear decoder that reconstructs the missing values. The loss is computed only on masked positions, encouraging the encoder to learn typical patterns of heat-pump electricity use under varying weather conditions without

relying on inspection labels.

In the lower branch, the pretrained encoder is reused via transfer learning and fine-tuned together with a small MLP head on the subset of households with inspection reports. The classification head outputs a probability distribution over the consumption categories (e.g., low, normal, and rather high) for each input window.

3.1. Data pre-processing

Algorithm 1 Data preprocessing and Household-level Splitting

Require: HEAPO tables; window length L ; stride S ; max missing ratio $r_{\max}$

Ensure: Disjoint Window sets $D_{\text{train}}, D_{\text{val}}, D_{\text{test}}$

```

1:  $H \leftarrow \text{JOINHOUSEHOLDS}(\text{households}, \text{meta\_data}, \text{protocols})$ 
2:  $H \leftarrow \text{FILTERVALID}(H)$ 
3: Split  $H$  randomly into mutually exclusive subsets:  $H_{\text{train}}, H_{\text{val}}, H_{\text{test}}$ 
4:  $D_{\text{train}} \leftarrow \emptyset, D_{\text{val}} \leftarrow \emptyset, D_{\text{test}} \leftarrow \emptyset$ 
5: for each  $\text{split} \in \{\text{train}, \text{val}, \text{test}\}$  do
6:   for  $h \in H_{\text{split}}$  do
7:      $X_h \leftarrow \text{ALIGNSERIES}(\text{SMART15MIN}(h), \text{WEATHERDAILY}(h))$ 
8:     for  $z \in \text{SLIDINGWINDOWS}(X_h, L, S)$  do
9:       if  $\text{MISSINGRATIO}(z) \leq r_{\max}$  then
10:         $D_{\text{split}} \leftarrow D_{\text{split}} \cup \{(z, h, c(h))\}$ 
11:       end if
12:     end for
13:   end for
14: end for
15: return  $D_{\text{train}}, D_{\text{val}}, D_{\text{test}}$ 

```

Alg. 1 summarizes how we transform the raw HEAPO tables into model-ready time-series windows. The procedure starts by building a unified household index: the `households`, `meta_data` and `protocols` tables are joined to obtain, for each household, its basic attributes, the associated weather station and, when available, the inspection label describing the heat-pump electricity consumption category. Households that lack a valid weather mapping or 15-minute smart-meter data are discarded at this stage.

For every remaining household, the algorithm then constructs a single aligned time series by merging 15-minute electricity measurements with the corresponding weather records. The weather variables are resampled to the same 15-minute grid and synchronized with the smart-meter timestamps, so that each time step contains both consumption and outdoor temperature information. From this aligned series, the procedure extracts overlapping sliding windows of fixed length. Windows with too many missing or invalid readings are removed, while the others are stored in a global pool D_{all} . If the originating household has an inspection label, the same window is also added to a labeled pool D_{lab} with its consumption category. In this way, the algorithm yields two coherent datasets: a large set of unlabeled windows for self-supervised pretraining and a smaller, fully labeled subset for downstream classification. To strictly prevent data leakage—a common pitfall in time-series forecasting where overlapping windows share data points—we enforce a *household-level split* prior to window extraction. As detailed in Alg. 1, the pool of valid households is first partitioned into mutually exclusive training, validation, and test sets. Sliding windows are subsequently generated exclusively within each subset. This guarantees that overlapping segments from the same household cannot inadvertently leak across different data splits during both the self-supervised pretraining and supervised fine-tuning phases.

3.2. Self-Supervised Learning

Figure 3 illustrates the self-supervised masked reconstruction task used to pretrain our encoder. We operate on 7-day windows of smart-meter electricity consumption and outdoor temperature with 15-minute resolution. Each window is represented as a sequence $\{x_t\}_{t=1}^L$, where $x_t \in \mathbb{R}^2$ contains the consumption and temperature at time step t .

Instead of standard random point masking, which is prone to trivial linear interpolation in highly autocorrelated series, we employ *temporal block masking*. We randomly select starting indices and mask contiguous chunks of time steps until the overall mask ratio ρ (here $\rho = 0.3$) is reached. Let $m_t \in \{0, 1\}$

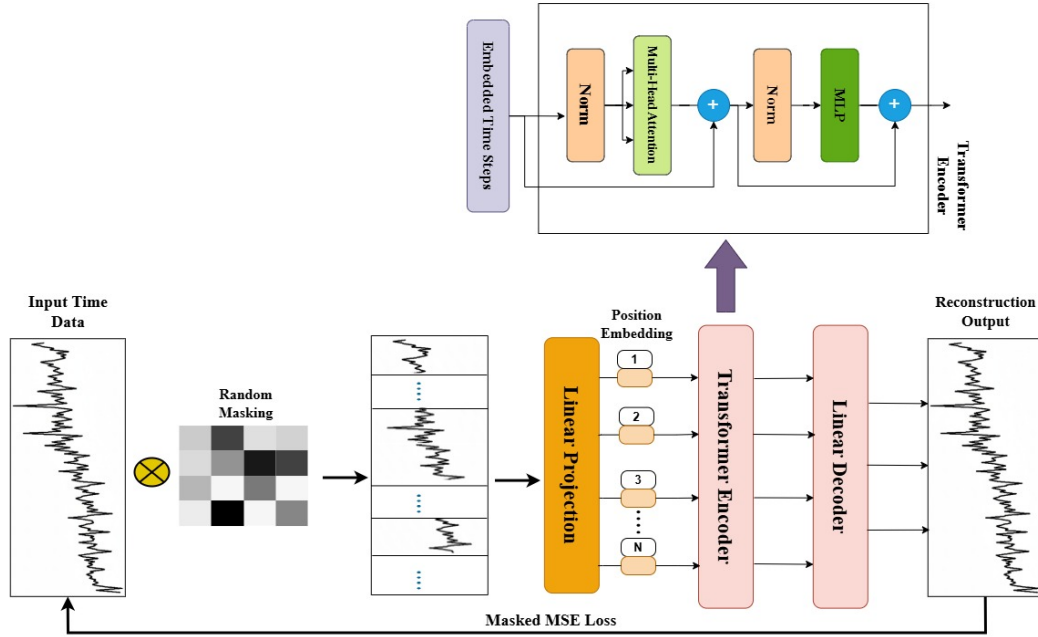


Figure 3. Self-supervised masked reconstruction on 7-day windows of smart-meter and weather data. A random subset of time steps is masked, the partially observed sequence is embedded and processed by a transformer encoder, and a linear decoder reconstructs the masked values. A masked mean-squared error goal is used to optimize the model.

denote the resulting binary mask, with $m_t = 1$ indicating that time step t is hidden within a block. The masked input sequence $\{\tilde{x}_t\}$ is obtained by

$$\tilde{x}_t = (1 - m_t) \odot x_t, \quad (1)$$

so that the model only observes unmasked values while still receiving positional information for all time steps.

Each masked vector $\tilde{x}_t$ is mapped to a d_{model} -dimensional embedding by a linear projection

$$z_t = W_{\text{in}} \tilde{x}_t + b_{\text{in}} \in R^{d_{\text{model}}}. \quad (2)$$

A positional encoding $p_t \in R^{d_{\text{model}}}$ is added to encode the position of each time step within the 7-day window:

$$e_t = z_t + p_t. \quad (3)$$

The embedded sequence $\{e_t\}_{t=1}^L$ is subsequently passed through a stack of transformer encoder layers that combine multi-head self-attention with feed-forward sublayers, producing a series of hidden representations $\{h_t\}_{t=1}^L$:

$$\{h_t\}_{t=1}^L = \text{TransformerEncoder}(\{e_t\}_{t=1}^L). \quad (4)$$

A lightweight linear decoder is applied independently at each time step to reconstruct the original two-dimensional signal:

$$\hat{x}_t = W_{\text{out}} h_t + b_{\text{out}} \in R^2. \quad (5)$$

The training is driven by a hidden mean square error loss, which only penalizes reconstruction errors at hidden locations:

$$L_{\text{mask}} = \frac{1}{\sum_{t=1}^L m_t} \sum_{t=1}^L m_t \|x_t - \hat{x}_t\|_2^2. \quad (6)$$

In general, by forcing the model to infer missing values from the surrounding context and weather information, the transformer encoder learns representations that capture typical patterns of heat pump operation without using any test labels. These pre-trained encoder weights are then reused in the downstream classification stage.

3.3. Fine-tuning Model

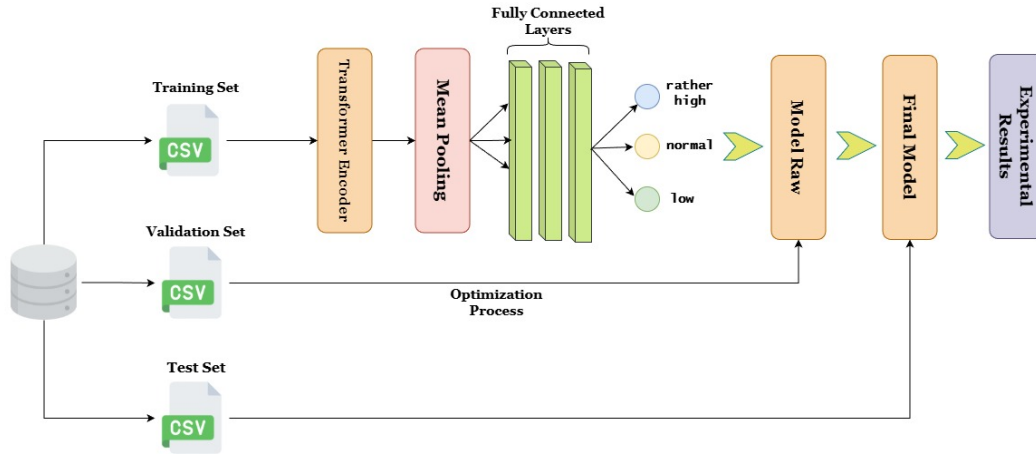


Figure 4. Fine-tuning the pipeline for three-level consumption classification. Labeled 7-day windows are split into training, validation, and test sets, passed through the pretrained transformer encoder with mean pooling, and classified by a small fully connected head into low, normal, or rather high.

Figure 4 shows the supervised fine-tuning stage built on top of the masked reconstruction encoder. We first restrict the data to households for which an inspection report provides a categorical assessment of annual electricity consumption, denoted by $y(h) \in \{\text{low}, \text{normal}, \text{rather high}\}$. Following the strict data isolation protocol established in the preprocessing stage, the labeled households are partitioned into mutually exclusive training, validation, and test splits prior to window generation. This guarantees that no temporal overlap contamination occurs between the splits, ensuring a rigorous evaluation of the model's generalization capabilities.

For a labeled window i we denote the 15-minute sequence by $\{x_{i,t}\}_{t=1}^L$ with $x_{i,t} \in \mathbb{R}^2$ containing electricity use and outdoor temperature. The sequence is processed by the pretrained transformer encoder, yielding hidden states $\{h_{i,t}\}_{t=1}^L$. We aggregate them by mean pooling

$$u_i = \frac{1}{L} \sum_{t=1}^L h_{i,t}, \quad (7)$$

and feed the pooled representation into a small fully connected classifier. Let

$$a_i = \sigma(W_1 u_i + b_1), \quad (8)$$

$$\ell_i = W_2 a_i + b_2, \quad (9)$$

where $\sigma(\cdot)$ is a non-linear activation and $\ell_i \in \mathbb{R}^3$ are the logits for the three classes. Class probabilities are obtained by a softmax

$$p_i = \text{softmax}(\ell_i). \quad (10)$$

The encoder and classifier are trained jointly with the standard cross-entropy loss over all labeled windows in a mini-batch,

$$L_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^N \log p_{i,y_i}, \quad (11)$$

where the ground-truth class of the home that generated the window i is indicated by y_i . The training set is used to optimize the model parameters, and the epoch that produces the best validation performance is chosen for assessment. Following training, the final model is applied to every window in every home in the test set. The experimental findings shown in Section 4.2. are calculated by adding the window-level predictions.

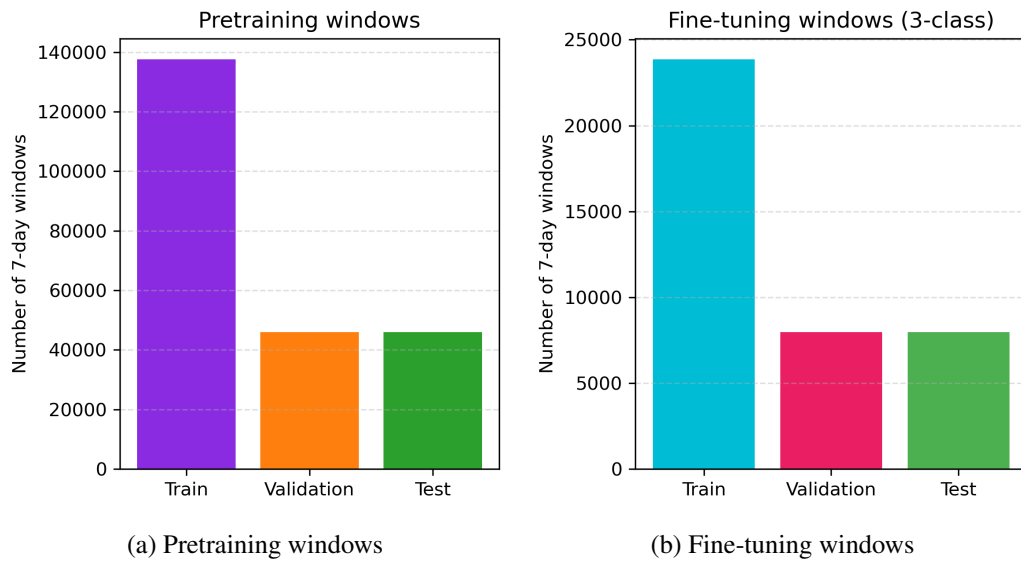


Figure 5. Number of 7-day windows allocated to the train, validation, and test sets in the two stages of our framework: (a) self-supervised pretraining on all households with 15-minute smart meter data; (b) supervised fine-tuning on the subset of households with inspection-based consumption labels.

4. EVALUATION

4.1. Experimental Settings

A workstation running Windows 11 Pro and equipped with 32 GB of RAM and an NVIDIA RTX 3060 (12 GB) GPU was used to conduct all experiments. The preprocessing, model implementation, and evaluation pipelines were developed in Python 3.10, utilizing PyTorch for deep learning models and Pandas for data manipulation.

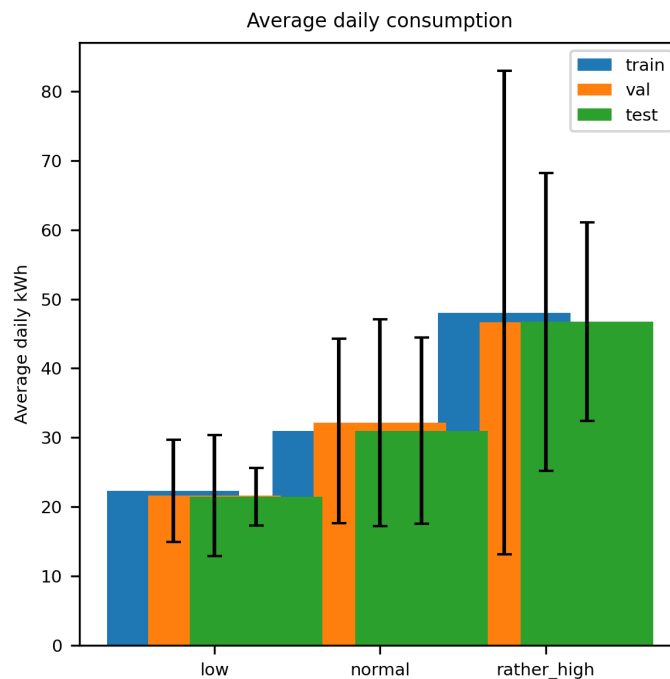


Figure 6. Average daily heat-pump use by class.

Fig. 5 illustrates the partitioning of the 7-day windows used in our experiments. For both the self-supervised pretraining stage (Fig. 5a) and the supervised fine-tuning stage (Fig. 5b), approximately 60% of the windows were assigned to the training set, while the remaining 40% were evenly divided between validation and testing. The pretraining data consists of windows from all households with 15-minute smart-meter readings, whereas the fine-tuning stage exclusively uses households for which test reports and three-level consumption labels are available. Fig. 6 illustrates the average daily heat pump electricity consumption across the low, normal, and rather high categories for the train, validation, and test splits. Across all partitions, the mean usage increases correspondingly from low to rather high. However, the substantial error bars indicate considerable intra-class variance, particularly within the 'rather high' group. The hyperparameter settings for all stages are summarized in Table 1.

Table 1. Main model and training hyperparameters for self-supervised pretraining and supervised fine-tuning.

Param.	Pretraining	Fine-tuning
Window length L	7 days (672)	7 days (672)
d_{model}	128	128
Encoder layers	4	4
Attention heads	4	4
Batch size	64	64
Learning rate	1×10^{-3}	5×10^{-4}
Epochs	50	40
Mask ratio	0.3	—
Loss	masked MSE	cross-entropy

Furthermore, to place our approach in context, we compare it against a set of widely used time-series baselines. As recurrent models, we include a vanilla recurrent neural network (RNN), a long short-term memory network (LSTM) and a gated recurrent unit (GRU), which all update a hidden state along the sequence and can capture short- to medium-range temporal dependencies. In addition, we consider a temporal convolutional network (TCN), which models sequences via stacks of dilated 1D convolutions, and TimesNet, a recent architecture that learns temporal patterns in a two-dimensional transformed representation of the input. Finally, we add TS2Vec as a self-supervised baseline that first pretrains a general-purpose time-series encoder and then trains a simple classifier on top of the resulting representations. The main architectural choices and training hyperparameters for these baseline models are summarised in Table 2.

Table 2. Main model and training hyperparameters for the baseline models.

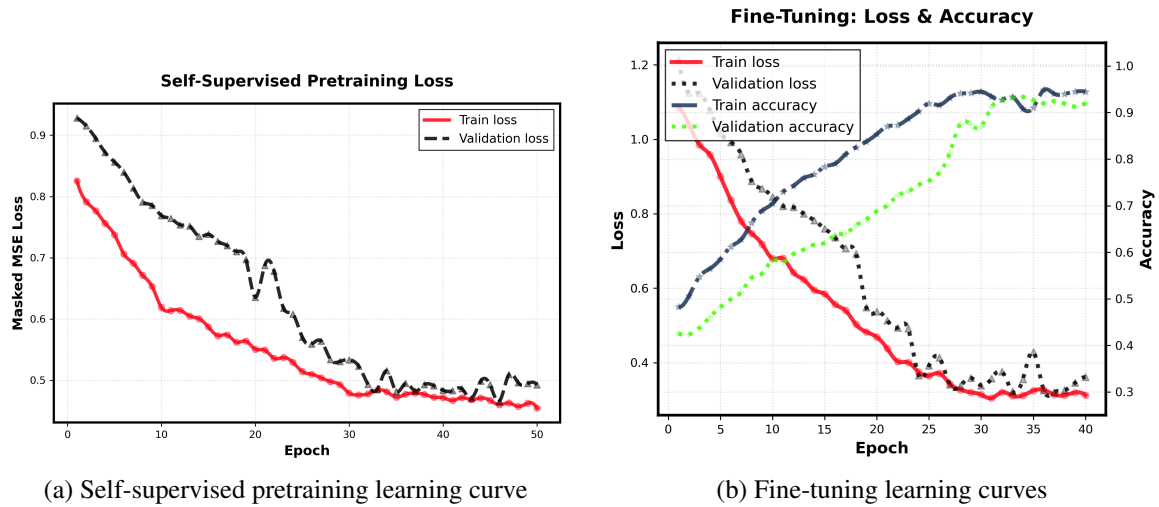
Model	Hidden / channels	Layers	Epochs	LR
RNN [32]	128 units	10	40	5×10^{-4}
LSTM [33]	128 units	10	40	5×10^{-4}
GRU [34]	128 units	10	40	5×10^{-4}
TCN [35]	64 channels	4 blocks	40	5×10^{-4}
TS2Vec [29]	128 dim	4 layers	40	5×10^{-4}
TimesNet [28]	128 channels	4 stacks	40	5×10^{-4}

4.2. Experimental Results

This section presents the results through visualizations, including the confusion matrix and the training performance graphs. Additionally, a comparative table is provided to summarize the Accuracy, F1-Score, and Recall metrics across the evaluated models.

First, we present Fig. 7a, which depicts the masked reconstruction loss in the self-supervised pretraining phase. Both the training and validation masked MSEs decrease rapidly in the first epochs, then decrease more slowly before reaching a stable threshold. The validation curve is still slightly higher than the training curve but has the same shape, indicating that the encoder learns consistent models of the heat pump power demand without any divergence between the training and validation data. This suggests that the self-supervised objective provides a reliable initialization for the subsequent supervised fine-tuning.

Second, we report the learning curves during supervised fine-tuning in Fig. 7b. The training loss remains nearly unchanged for the first 40 epochs, and the validation loss follows a similar pattern before leveling off in the later stages. Meanwhile, both training and validation accuracies increase steadily from around 0.5 to over 0.9, with the validation curve consistently lagging slightly behind the training curve. The narrow gap between the two curves in the final epochs indicates only minor overfitting and confirms that the pre-trained encoder can be effectively adapted to the three-class prediction task even with a relatively small labeled dataset.



(a) Self-supervised pretraining learning curve

(b) Fine-tuning learning curves

Figure 7. Learning curves for the proposed framework: (a) self-supervised pretraining with masked reconstruction loss on the training and validation windows, and (b) fine-tuning with cross-entropy loss and classification accuracy on the training and validation sets.

Simultaneously, Fig. 8 illustrates the training and validation loss and accuracy curves for the baseline sequence models on the three-class HEAPO household classification task over 40 epochs. Specifically, subfigures (a)-(c) present the performance of the RNN, LSTM, and GRU models, respectively, while subfigure (d) reports the results obtained with TimesNet. In addition, subfigure (e) depicts the training dynamics of the TCN baseline, and subfigure (f) shows the results achieved by TS2Vec. Across all configurations, the baseline models exhibit a consistent reduction in training loss and an improvement in accuracy. However, pronounced gaps between the training and validation curves are observed, particularly for the RNN, LSTM, and GRU, indicating limited generalization capabilities. More recent architectures, including TimesNet, TCN, and TS2Vec, demonstrate comparatively more stable validation performance; nevertheless, noticeable fluctuations remain. This highlights the inherent challenges of learning discriminative representations under limited supervision in real-world energy consumption data.

Third, we present the confusion matrix in Fig. 9. The rows represent the true classes, and the columns represent the predicted classes. The model correctly identifies low, medium, and rather high consumption at approximately 92.3%, 94.3%, and 91.9% of households, respectively, while off-diagonal entries remain below 10%. Most of the remaining errors are slight confusions between neighboring groups (e.g., normal vs. rather high). These misclassifications likely stem from external factors not fully captured by the inputs, such as unusual weather anomalies, undocumented use of supplementary electric heaters, or variations in building envelope insulation quality. Overall, the matrix shows that the proposed model can reliably separate the three risk levels across all classes.

Finally, Table 3 reports the test performance of the proposed model and all time-series baselines on the three-class household classification task. For each architecture, we run the training procedure five times with different random seeds (16,32,64,120,150) and report mean $\pm$ standard deviation over these runs. The recurrent baselines (RNN, LSTM, GRU) and the more advanced models (TCN, TimesNet, TS2Vec) obtain accuracies in the range of roughly 0.84–0.90, with F1-score and recall at similar levels and standard deviations on the order of a few percentage points. In contrast, the proposed self-supervised transformer reaches about 0.93 ± 0.01 test accuracy and around 0.91 ± 0.01 for both F1-score and recall, indicating a consistent margin over the baselines that exceeds the variability across random seeds. This suggests that the pretrained encoder extracts more informative representations from the smart-meter and weather time series than the purely supervised sequence models. Additionally, we ran a few additional tests to verify our design choices. A Wilcoxon signed-rank test confirmed that the accuracy improvement over the TS2Vec baseline is statistically significant ($p < 0.05$). Regarding the input features, training the model solely on the 15-minute load without the daily outdoor temperature caused the test accuracy to drop to about 88%. This confirms that weather context is essential to

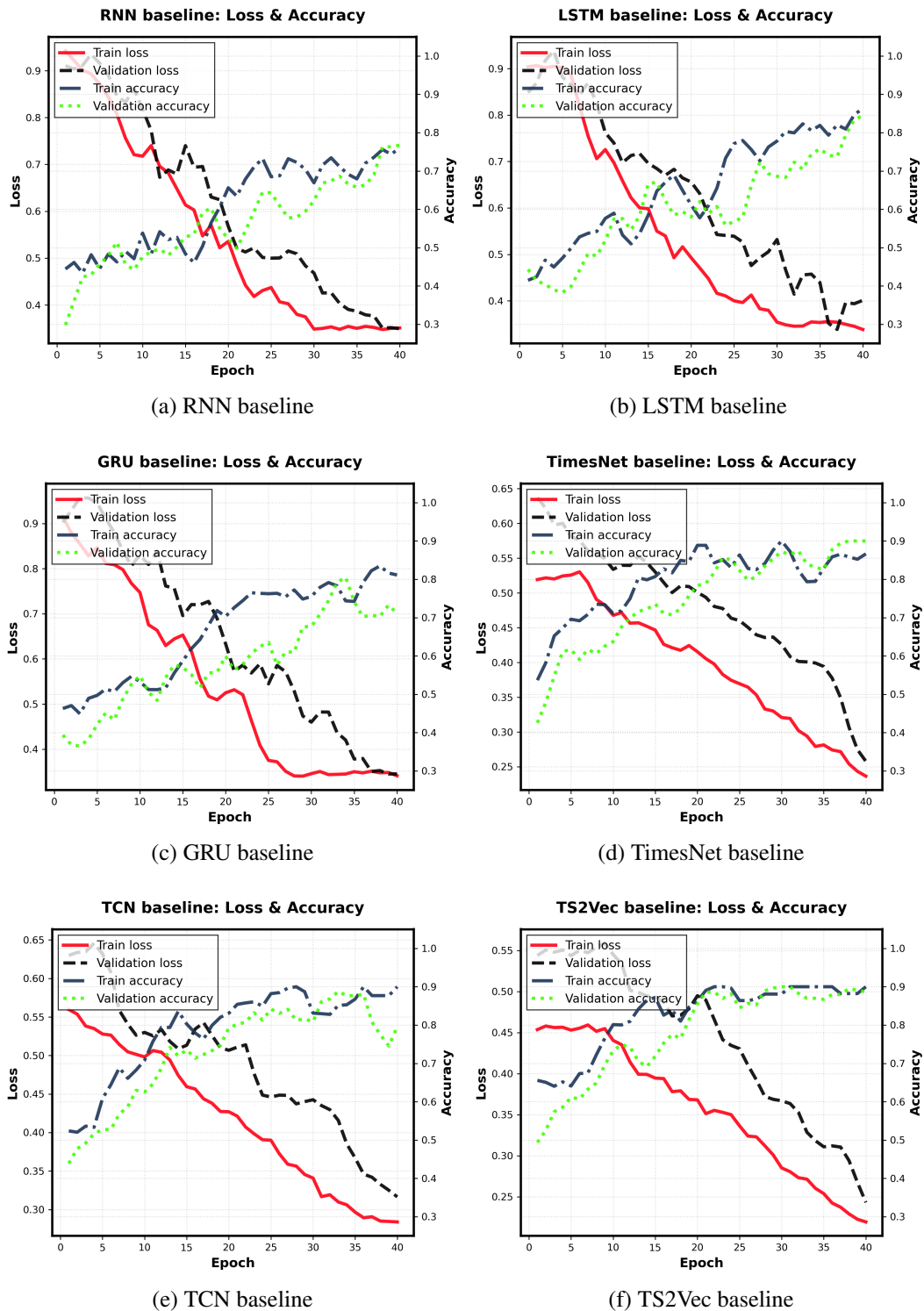


Figure 8. Training and validation loss/accuracy curves for the baseline sequence models on the HEAPO three-class classification task. Each panel shows cross-entropy loss (left axis) and accuracy (right axis) over 40 epochs for (a) RNN, (b) LSTM, (c) GRU, (d) TimesNet, (e) TCN, and (f) TS2Vec.

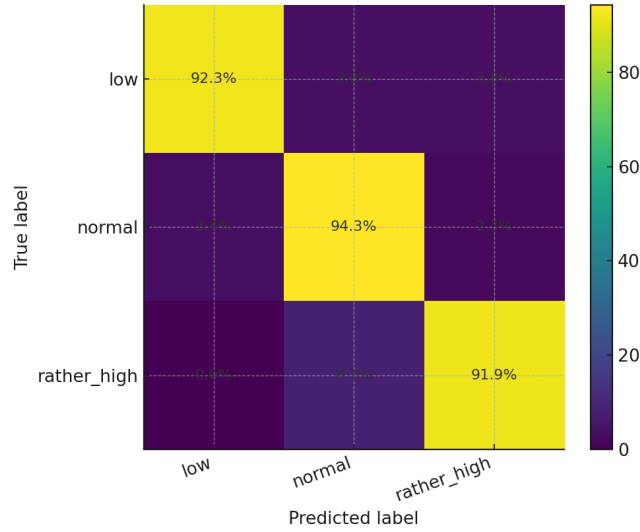


Figure 9. Confusion matrix of the three-class heat-pump consumption classifier on the test households.

distinguish between naturally high demand on cold days and actual system inefficiencies. Furthermore, because smart meter data is highly autocorrelated, we used temporal block masking instead of random point masking to prevent the model from learning trivial linear interpolations. Testing various ratios showed that $p = 0.3$ offers the best trade-off for representation learning, whereas higher masking ratios (e.g., 0.5) hide too much operational cycle information and degrade downstream performance.

Table 3. Performance comparison (mean $\pm$ standard deviation over five runs) between the proposed method and baseline models.

Model	Accuracy	F1 score	Recall
RNN [32]	0.84 ($\pm$ 0.02)	0.81 ($\pm$ 0.02)	0.80 ($\pm$ 0.03)
LSTM [33]	0.86 ($\pm$ 0.02)	0.84 ($\pm$ 0.02)	0.83 ($\pm$ 0.02)
GRU [34]	0.88 ($\pm$ 0.01)	0.86 ($\pm$ 0.02)	0.85 ($\pm$ 0.02)
TCN [35]	0.88 ($\pm$ 0.01)	0.86 ($\pm$ 0.02)	0.85 ($\pm$ 0.02)
TimesNet [28]	0.89 ($\pm$ 0.01)	0.87 ($\pm$ 0.01)	0.86 ($\pm$ 0.02)
TS2Vec [29]	0.90 ($\pm$ 0.01)	0.88 ($\pm$ 0.01)	0.87 ($\pm$ 0.02)
This study	0.93 ($\pm$ 0.01)	0.91 ($\pm$ 0.01)	0.91 ($\pm$ 0.01)

5. CONCLUSIONS

In this paper, we investigated whether routinely collected smart-meter and weather data can support the pre-selection of heat-pump installations for on-site inspections. Using the HEAPO dataset, which links 15-minute electricity demand, daily outdoor temperature and expert inspection reports at household level, we proposed a transformer-based framework with self-supervised pretraining. A masked reconstruction objective was first used to train an encoder on multi-day windows from largely unlabeled households, after which the encoder was fine-tuned with a lightweight classification head on the limited subset of homes with inspection labels. The resulting model achieves high test accuracy and reliably distinguishes between *low*, *normal* and *rather high* consumption categories, clearly outperforming non-pretrained transformers and simple feature-based baselines.

From a practical viewpoint, our results suggest that utilities and service companies can exploit existing smart-meter infrastructures to focus expert attention on households with a higher probability of inefficient heat-pump operation, potentially reducing unnecessary visits and improving the effectiveness of field work. At the same time, the study is constrained by the subjective nature of the inspection-based labels and by the specific geographic and regulatory context of HEAPO, so retraining or calibration would be required before deployment in other regions.

Future work will address these limitations in several directions. First, we plan to evaluate the approach on additional datasets and under different tariff schemes to assess robustness and transferability. Second, while

our current study comprehensively benchmarks against state-of-the-art deep learning sequence models, we did not compare the proposed approach against traditional engineering baselines, such as weather-adjusted heating curve residual analysis or degree-day normalized linear regression. Additionally, due to computational constraints, this study did not benchmark against dedicated patch-level SSL architectures (e.g., PatchTST) or perform large-scale ablations on the pre-training dataset size. Future research should incorporate these physical and advanced SSL baselines, investigate the impact of pre-training data scale, and include richer contextual information, such as building thermal inertia, renovation status, and PV generation, to provide a more holistic evaluation. Finally, extending the framework with better uncertainty quantification could make the recommendations more transparent for both engineers and policy makers.

REFERENCES

- [1] International Renewable Energy Agency (IRENA), "Power to heat and cooling: Status," <https://www.irena.org/Innovation-landscape-for-smart-electrification/Power-to-heat-and-cooling/Status>, 2023.
- [2] International Energy Agency (IEA), "Buildings," <https://www.iea.org/energy-system/buildings>, 2025, accessed 2025-12-03.
- [3] O. Usman, P. T. Iorember, O. Ozkan, and A. A. Alola, "The asymmetric effect of household and commercial energy efficiency on u.s. energy-related co2 emissions," *Innovation and Green Development*, vol. 4, no. 5, p. 100293, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2949753125000906>
- [4] T. Brudermueller, E. Fleisch, M. González Vayá, and T. Staake, "Heapo – an open dataset for heat pump optimization with smart electricity meter data and on-site inspection protocols," 2025. [Online]. Available: <https://arxiv.org/abs/2503.16993>
- [5] N. Singhal, A. Kadam, P. Kumar, H. Singh, A. Thakur, and Pranay, "Study of recent image restoration techniques: A comprehensive survey," *Jordanian Journal of Computers and Information Technology (JJCIT)*, vol. 11, no. 02, pp. 211 – 237, 2025.
- [6] L. A. Qadi and A. E. Hozayfa El Rifai, Safa Obaid, "A scalable shallow learning approach for tagging arabic news articles," *Jordanian Journal of Computers and Information Technology (JJCIT)*, vol. 06, no. 03, pp. 263 – 280, 2020.
- [7] C. N. Anh, T. B. Thao, T. B. Hung, T. T. Huong, and N. V. Hung, "Optimizing transportation costs: enhancing logistics efficiency and resource utilization in dynamic environments," *Informatics and Automation*, vol. 24, no. 3, pp. 856–883, 2025.
- [8] N. Viet Hung, D. L. Luu, Q. T. Pham, and C. Lupu, "Comparative analysis of different spacing policies for longitudinal control in vehicle platooning," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 239, no. 10-11, pp. 4810–4825, 2025.
- [9] N. A. Cao, T. B. Thao, N. S. Dong, P. H. Truyen, and N. V. Hung, "Coordinated start and automated vehicle platoons for arterial intersection improvement to solve the capacitated vehicle routing problem," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 25 978–25 983, 2025.
- [10] D. L. Luu, N. V. Hung, and C. Lupu, "Tracking trajectory by using the polynomial method for acc system based on smart car platform," in *2023 17th International Conference on Engineering of Modern Electric Systems (EMES)*. IEEE, 2023, pp. 1–4.
- [11] N. V. Hung, T. T. Huong, N. Tan, T. C. Doan, and N. Nam-Hoang, "Machine learning applications for delivery time prediction and freight planning," *Informatics and Automation*, vol. 24, no. 5, pp. 1379–1407, 2025.
- [12] T. Brudermueller, "Heapo – an open dataset for heat pump optimization," <https://github.com/tbrumue/heapo>, 2025, accessed 2025-12-03.

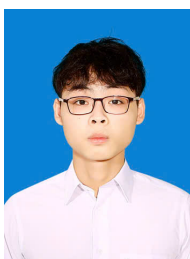
- [13] Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali, and A. Amira, "Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives," *Applied Energy*, vol. 287, p. 116601, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0306261921001409>
- [14] H. Rashid, N. Batra, and P. Singh, "Rimor: Towards identifying anomalous appliances in buildings," in *Proceedings of the 5th ACM International Conference on Systems for Built Environments (BuildSys)*, 2018, pp. 1–10.
- [15] K. Zhang, Q. Wen, C. Zhang, R. Cai, M. Jin, Y. Liu, J. Zhang, Y. Liang, G. Pang, D. Song, and S. Pan, "Self-supervised learning for time series analysis: Taxonomy, progress, and prospects," 2024. [Online]. Available: <https://arxiv.org/abs/2306.10125>
- [16] T. Brudermueller, U. Potthoff, E. Fleisch, F. Wortmann, and T. Staake, "Estimation of energy efficiency of heat pumps in residential buildings using real operation data," *Nature Communications*, vol. 16, 03 2025.
- [17] M. Jung, S. Jeong, S. Kim, D. Ku, K. Seo, J. Song, S. I. Na, G. Lee, and M. Kim, "Fault intensity variation of a heat pump system by operating temperature change," *Applied Thermal Engineering*, vol. 277, p. 127119, 06 2025.
- [18] B. Hofer, A. Heinz, and F. Wotawa, "Detecting soft faults in residential heat pumps via regression-based residual analysis," *Sustainable Cities and Society: Advances*, vol. 2, no. 2, p. 100059, 2026. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S3051052X26000379>
- [19] A. Weigert, "Identification and classification of heat pump problems in the field and their implication for a user-centric problem recognition," *Energy Informatics*, vol. 5, 12 2022.
- [20] Y. Guo, Y. Liu, Z. Jin, Y. Wang, C. Du, C. Liu, and W. Li, "Fault diagnosis of heat pump system with different transfer tasks based on cross domain deep transfer learning," *International Journal of Refrigeration*, vol. 181, 10 2025.
- [21] R. Alijani and F. Leonforte, "Performance analysis of multi-source heat pumps: A regression-based approach to energy performance estimation," *Sustainability*, vol. 17, p. 6804, 07 2025.
- [22] T. Wilcox, N. Jin, P. Flach, and J. Thumim, "A big data platform for smart meter data analytics," *Computers in Industry*, vol. 105, pp. 250–259, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0166361518303749>
- [23] Y. Wang, Q. Chen, C. Kang, Q. Xia, and M. Luo, "Sparse and redundant representation-based smart meter data compression and pattern extraction," *IEEE Transactions on Power Systems*, vol. 32, pp. 2142–2151, 05 2017.
- [24] B. Amangeldy, T. Imankulov, N. Tasmurzayev, G. Dikhanbayeva, and Y. Nurakhov, "A review of artificial intelligence and deep learning approaches for resource management in smart buildings," *Buildings*, vol. 15, 07 2025.
- [25] D. Fährmann, L. Martín, L. Sánchez, and N. Damer, "Anomaly detection in smart environments: A comprehensive survey," *IEEE access*, vol. 12, pp. 64 006–64 049, 2024.
- [26] Y. Song, S. Kuang, J. Huang, and D. Zhang, "Unsupervised anomaly detection of industrial building energy consumption," *Energy and Built Environment*, vol. 7, no. 2, pp. 351–363, 2026. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666123324001144>
- [27] Z. Zhang, Y. Chen, H. Wang, Q. Fu, J. Chen, and Y. Lu, "Anomaly detection method for building energy consumption in multivariate time series based on graph attention mechanism," *PLOS ONE*, vol. 18, no. 6, pp. 1–23, 06 2023. [Online]. Available: <https://doi.org/10.1371/journal.pone.0286770>
- [28] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "Timesnet: Temporal 2d-variation modeling for general time series analysis," 2023. [Online]. Available: <https://arxiv.org/abs/2210.02186>

- [29] Z. Yue, Y. Wang, J. Duan, T. Yang, C. Huang, Y. Tong, and B. Xu, "Ts2vec: Towards universal representation of time series," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 8, 2022, pp. 8980–8987.
- [30] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly transformer: Time series anomaly detection with association discrepancy," *arXiv preprint arXiv:2110.02642*, 2021.
- [31] E. Eldele, M. Ragab, Z. Chen, M. Wu, C.-K. Kwoh, X. Li, and C. Guan, "Self-supervised contrastive representation learning for semi-supervised time-series classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 15 604–15 618, 2023.
- [32] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.
- [33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [34] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.
- [35] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

BIOGRAPHIES OF AUTHORS



Hoang Thi Thu   received the B.E. degree from the Posts and Telecommunications Institute of Technology (PTIT), Vietnam, in 2017, and the M.E. degree from the Posts and Telecommunications Institute of Technology (PTIT) in 2019. She is currently working at the Telecommunications Department 1, Posts and Telecommunications Institute of Technology, Vietnam. Her research interests include the Internet of Things (IoT), wireless sensor networks (WSNs), telecommunications networks, edge computing, and emerging networking technologies. Her current research focuses on the development and application of intelligent and efficient networking solutions for resource-constrained and distributed systems. She can be reached via email at thuht@ptit.edu.vn.



Phi Dinh Huynh   is an undergraduate student at the Faculty of Information Technology, East Asia University of Technology, and a Research Assistant at the Mobilab Laboratory since 2023. His research focuses on computer networking, computer vision, data mining, and intelligent systems. He is particularly interested in integrating autonomous vehicles with AI-driven technologies to enhance smart mobility and intelligent communication systems. With strong enthusiasm for research, he continues to explore the synergy between machine learning, computer vision, and networked systems. He can be contacted at email: 20222072@eaut.edu.vn.



Ta Trung Kien   received the Bachelor's degree in Labor Relations from Trade Union University, Vietnam, in 2018, and the Master's degree in Commercial Marketing from Thuongmai University, Vietnam, in 2023. He is currently a lecturer at the Faculty of Business Administration and also serves as the Director of Admissions and Marketing at the International Training and Cooperation Institute (ITCI), East Asia University of Technology (EAUT), Vietnam. His research interests include business administration, marketing management, information technology, artificial intelligence, and related applications. He can be contacted at email: kientt@eaut.edu.vn.



Le Mai Nam   received the Bachelor's degree in Applied Mathematics and Informatics from the University of Natural Sciences, Vietnam National University, Hanoi, in 2007, and his Master's degree in Information Systems from the Military Technical Academy in 2015. His research interests include data structures and algorithms, optimization methods, and mobile application development. He has contributed to various projects and academic activities in the fields of computer science and software engineering, focusing on practical applications of computational optimization and mobile technologies in education and industry. He can be contacted at email: namlm@eaut.edu.vn.



Le Hong Duy   received the Bachelor's degree in Electronics and Telecommunications from Hanoi University of Science and Technology, Vietnam, in 2016, and the Master's degree in Information Technology from Hanoi University of Science and Technology, Vietnam, in 2025. He is currently a lecturer at CMC University, Vietnam. His research interests include Generative Adversarial Networks (GANs), biometric image recognition, artificial intelligence, and computer vision. He can be contacted at email: lhdud@cmcu.edu.vn.



Pham Hong Thinh   received the B.E. and M.E. degrees in Electronics and Telecommunications from the Hanoi University of Technology, Vietnam, in 2002 and 2006, respectively. He obtained his Ph.D. degree in Telecommunications Engineering from the Hanoi University of Science and Technology, Vietnam, in 2021. Since 2002, he has been working at Quy Nhon University, Vietnam. His research interests include Quality of Experience (QoE) optimization, Software-Defined Networking (SDN), artificial intelligence, machine learning, Virtual Reality (VR), and content adaptation over HTTP. He can be reached via email at phamhongthinh@qnu.edu.vn



Nguyen Viet Hung   received the B.Sc. degree in Informatics Pedagogy from the Faculty of Engineering Technology, Ha Tinh University, Vietnam, in 2012, and the M.Sc. degree in Information Technology from the Faculty of Information Technology, Ho Chi Minh City University of Technology, Vietnam, in 2016. He is currently pursuing the Ph.D. degree in Telecommunications Engineering at Hanoi University of Science and Technology, Vietnam, in 2025. He is the Director of Information Technology Programs at the International Training and Cooperation Institute (ITCI), East Asia University of Technology. His research interests include multimedia communications, network security, artificial intelligence, traffic engineering in next-generation networks, machine learning, deep learning, computer vision, data and web mining, autonomous vehicles, QoE/QoS assurance for network services, green networking, and related applications. He can be contacted at email: hungnv@eaut.edu.vn.