19. Symposium Energieinnovation, 11.–13.02.2026, Graz/Österreich

Comparing Load-Forecasts of Residential Heatpumps with Transformer and XGB on Field Data

Marcel Lüdecke[1]* , Timon Justi[1], Michel Meinert[1]* , Bernd Engel[1]

[1]*Technische Universität Braunschweig, elenia Institute for High Voltage Technology and Power Systems, Braunschweig, Germany*
[*]*m.luedecke@tu-braunschweig.de*

Keywords: Forecasting, Heat Pump, Prosumer, Energy Management, Transformer, XGB, Forecasting Error, Model

Abstract

Accurate device-level forecasting of residential heat pump power consumption is a key enabler for advanced Prosumer energy management, yet forecasting performance is often limited by user-driven variability and incomplete measurement data. This work examines day-ahead forecasting of individual air-source heat pumps at 15-minute resolution using a large field dataset from 308 devices, combining lagged load, weather, and calendar features. A globally trained XGB model, locally trained XGB and Transformer models, and benchmark methods (linear regression and 24 h persistence) are comparatively evaluated with respect to R^2, normalized $RMSE$, and a peak error metric. Results reveal pronounced performance heterogeneity across devices, with the global XGB model achieving $R^2 > 0.8$ and $nRMSE < 0.05$ for series with regular daily peaks, but negative R^2 and large peak errors for highly irregular time series. Transformers do not consistently outperform XGB and tend to overfit to noisy data despite considerable model capacity and training effort. Feature ablation experiments identify lagged load values as the dominant predictors and indicate that temperature and periodic features alone yield poor forecasts. Generalization analyses show that models trained on time series with consistent patterns transfer reasonably well to unseen, regular time series, whereas models calibrated on irregular time series generalize poorly and are sometimes inferior to persistence, highlighting the central role of the intrinsic load structure in forecastability. The findings underscore that, in the studied setting, robust, computationally efficient tree-based ensembles remain competitive with deep learning methods.

1 Introduction

1.1 Motivation

To meet the targets set by the German federal government, greenhouse gas emissions in the building sector must be drastically reduced in the coming years [1]. One way to achieve this goal is to widely install heat pumps (HPs) for space heating and domestic hot water supply. During periods of high heat demand, photovoltaic (PV) generation in Germany and at comparable latitudes is comparatively low [2]. Nevertheless, accurate HP load forecasting supports the efficient utilization of PV energy and, importantly, helps to avoid overloading the grid connection point [3]. In particular, day-ahead forecasts with 15 min resolution provide the temporal detail required to anticipate load peaks and align HP operation with on-site PV production and tariff signals. Such forecasts form a key input to advanced residential energy management systems, which rely on them to coordinate flexible demand, enhance PV self-consumption, and maintain local grid stability [3, 4].

1.2 Contribution

This work addresses day-ahead forecasting of residential Prosumer-level air-source heat pumps' (ASHPs) electricity demand at 15-min resolution, where user-driven variability strongly affects load profiles and complicates data-driven modeling. The study focuses on forecasting electrical demand using publicly available datasets and compares Transformer-based sequence models with eXtreme Gradient Boosting (XGB) in a high-resolution, day-ahead setting. Due to the strong influence of individual load patterns, the time series are classified into irregular and regular consumption types, and the impact of these types on forecasting accuracy is systematically analyzed. The chosen forecast horizon and 15-min resolution mirror typical requirements of residential energy management systems, enabling direct use of the forecasts for operational scheduling of HPs and PV-coupled Prosumer systems.

1.3 Paper Structure

The paper is structured as follows: Section 2 describes the related work. Section 3 describes the literature-based selection of the applied forecasting architecture. Section 4 describes the input data, its preprocessing, and the feature selection, as well as the implementation of the forecasting models. Section 5 analyzes the forecast accuracy. Section 6 discusses the limitations encountered and suggests possibilities for future work. Section 7 summarizes the essential findings and conclusions drawn from the research.

2 Related Work

Research on forecasting the electrical load of heat pumps spans a range of building types, system configurations, and modeling objectives. Several studies have analyzed ground-source or

air-to-water heat pumps in non-residential buildings such as laboratories [5], office buildings [6], or university facilities [7], often with the primary goal of evaluating control strategies or assessing system efficiency rather than providing operational day-ahead forecasts for Prosumer-level energy management. The work of Song [4] benefits from detailed sensor networks and access to internal system states, leading to high predictive performance but limiting their applicability to less-detailed residential environments.

Within the residential domain, research has increasingly examined aggregated electricity demand from heat pumps in multi-family houses [8] or at the energy community scale [9]. Semmelmann et al. [9] study day-ahead forecasting of aggregated heat pump and household loads using Random Forest (RF), XGB, Long Short-Term Memory (LSTM), and Transformer models and report that tree-based models perform best for household demand, while the Transformer model yields the lowest errors for aggregated heat pump loads. Other recent contributions propose hybrid or physics-informed models to estimate heat pump electricity load profiles [5, 10], generally focusing on aggregated behaviour [8, 9] and on quantifying the grid impacts of large-scale heat pump deployment [4, 11].

By contrast, there are comparatively few works on forecasting individual device-level electrical loads, despite their relevance for fine-grained demand response and appliance-level control. Appliance-level studies, such as those by Ji et al. [12] and Razghandi et al. [13], propose probabilistic models and deep learning architectures for forecasting the loads of individual household appliances, showing that device-specific consumption is highly influenced by user behaviour and exhibits considerable uncertainty [14, 15]. These findings support the observation that forecasting individual devices, including residential heat pumps, is more challenging than forecasting aggregated loads because of the dominance of idiosyncratic usage patterns.

To the current knowledge, no prior work has systematically investigated day-ahead forecasting of the electrical demand of individual residential heat pumps in single- and two-family homes using high-resolution (15 min) data, and directly comparing state-of-the-art tree-based methods and Transformer architectures [5, 9]. Existing studies either operate at higher aggregation levels, rely on additional system measurements that are rarely available in typical Prosumer settings, or pursue different objectives such as estimating heat demand rather than electrical load [8, 9]. The present work addresses this gap by focusing on individual residential HPs with limited measurement infrastructure and by explicitly evaluating XGB and Transformer on Prosumer-level field data, identifying forecasting constraints [9, 13, 14].

3 Selection of the Forecasting Method

In recent years, several literature reviews have provided structured overviews of state-of-the-art load forecasting techniques, including statistical, machine-learning, and deep-learning approaches [16–19, 47]. One common way to categorize forecasting methods is to distinguish between linear models and machine learning models [16], with deep learning forming a subfield of the latter [18]. Across these reviews, frequently used methods include linear regression, AutoRegressive (Integrated) Moving Average (ARMA/ARIMA), k-nearest neighbours (KNN), support vector machines (SVM), tree-based ensemble models such as RF and Gradient Boosted Trees, as well as a wide range of Artificial Neural Network (ANN) architectures, including FeedForward ANN (FFANN), Recurrent NN (RNN), LSTM, Convolutional NN (CNNs), and Transformer-based models. In recent years, deep-learning models have emerged as the most prominent techniques for short-term load forecasting, particularly when rich feature sets and large datasets are available [18].

However, deep learning methods are not universally applicable to all load forecasting problems. The suitability of a specific model family depends on application-specific factors such as data availability and quality, the temporal resolution and forecast horizon, the degree of nonlinearity in the underlying process, and the required level of interpretability [14]. Building-level and device-level loads often exhibit highly heterogeneous consumption patterns, implying that no single model class is appropriate for every use case. Moreover, the availability of explanatory variables (e.g., weather, occupancy, operational signals) and the granularity of metering strongly influence model choice and feature engineering requirements [21]. Consequently, recent reviews recommend selecting forecasting methods with respect to the forecasting task and its constraints in mind rather than applying deep learning models by default [21].

Traditional linear methods, such as linear regression (LinReg) and time-series models like ARMA and ARIMA, achieved early successes in load forecasting and remain widely used as reference, baseline, or benchmark models due to their simple architecture and good interpretability [14, 18, 22]. These models are computationally efficient and can perform competitively in settings with limited data, or where the load dynamics are approximately linear [47]. Nonetheless, their ability to capture complex, nonlinear interactions and long-range temporal dependencies is limited compared to more flexible machine learning approaches.

Among machine learning methods, KNN, SVM, and tree-based ensembles are popular choices for load forecasting, particularly when the problem can be formulated in a tabular feature space [14, 18]. Tree-based models such as RF and XGB are especially attractive due to their robustness to heterogeneous features, high computational efficiency, and strong predictive performance with relatively modest tuning effort [23]. These models have been successfully applied in numerous load forecasting studies at the household, building, and community levels and often serve as strong benchmarks against which more complex deep learning models are evaluated.

ANNs, inspired by biological neurons, process information through multiple layers of interconnected nodes, where the layer configuration determines the network architecture [16]. Popular ANN types in load forecasting include FFANNs, RNNs, LSTMs, and Gated Recurrent Units (GRUs) [16, 18],

which are particularly suited to capturing nonlinear relationships and temporal dependencies in energy time series [13, 21]. More recently, Transformer-based models have gained attention for time series forecasting. The Transformer architecture was first introduced in 2017 by [24] and has quickly become the most successful deep learning approach for natural language processing and computer vision [25]. In recent years, several variants of the original Transformer architecture have been developed for time series forecasting [10]. Transformers are gaining popularity in time series analysis due to their advantages over RNN-based models for sequence processing [25], further motivating their consideration alongside strong tree-based architectures such as XGB in this work.

4 Forecast of Electrical Load of Individual Heatpumps

A critical aspect of time series forecasting lies in the selection and preparation of the data used for model training. In the following, the data sources utilized in this study and the procedures for data preparation and cleaning will be presented. Subsequently, the feature preparation and selection process is described. Finally, the chosen models are introduced, and the training process is presented.

4.1 Data Selection and Preprocessing

As simulation data represent real-world conditions only to a limited extent [4], this work uses field data from the *Electrification of Heat Demonstration Project* [26]. The dataset comprises measurements from 306 ASHPs installed in residential buildings in the UK. The systems' rated power ranges from 5.0 to 16 kW and were monitored from January 2021 to September 2023. In addition, data from two ASHPs installed in single-family houses in Konstanz, Germany, are included. These data are obtained from the *Household Data dataset* provided by *Open Power System Data* [28] and cover a measurement period from May 2015 to March 2017. Overall, the available time series range from a few months to 32 months, with most spanning more than 2 years. For further processing, timestamps, accumulated electrical energy consumption (kWh), and outdoor air temperature (where available) are used. Electrical energy consumption is converted into load values (kW) and aggregated to 15-minute averages. During data preprocessing, missing values are linearly interpolated, and extreme outliers, caused by measurement errors that occur after time jumps in the measurement log, as well as time series segments with more than 3 consecutive missing measurements, are removed.

The data analysis comprises visual inspection, autocorrelation function (ACF) analysis, partial autocorrelation function (PACF) analysis, and frequency spectrum analysis. In general, the load time series of individual ASHPs do not exhibit a uniform or consistent pattern. This variability can be attributed to the wide range of application contexts and user behaviors under which heat pumps operate in real-world conditions, consistent with findings reported in the literature [15]. All load time series exhibit varying degrees of correlation between the load and its lagged values, as well as differing strengths of periodic behavior. The analysis further indicates that heat pumps operate in three distinct states [12], frequently switching between standby, medium, and high-load states. Medium load levels occur predominantly during the heating season. Load peaks of consistent magnitude are observed throughout the year, suggesting that medium load levels are primarily associated with space heating, while higher peak loads are likely related to domestic hot water production. However, due to the lack of detailed information on individual usage patterns, these interpretations cannot be conclusively validated.

4.2 Feature Engineering

To improve forecasting accuracy and to support the model in identifying patterns, additional features are added to the dataset [14, 18]. Typical features for short-term load forecasting (STLF) in the residential sector include weather data [32], building characteristics [39], occupancy patterns [40], calendar variables [41], historical load values [14], and cyclic calendar characteristics [47]. To reduce model complexity, only important features are used in the final forecast model [42]. To identify the most relevant features, a two-stage process is designed, following the approach proposed by [9]. This process involves collecting relevant features and selecting the most important ones using a random forest–based feature-importance analysis and Spearman correlation. The Spearman coefficient measures the monotonic relationship between two variables by calculating the linear correlation between their ranks. Based on a ranking of the results, features with sufficiently strong correlations are identified [43]. The Random Forest algorithm is an ensemble method for identifying related variables and capturing complex interactions [42]. It provides a ranking that indicates the percentage contribution of each feature to improving the forecast [44]. Irrelevant features are excluded from the final forecasting model.

Additional weather data (relative humidity, wind speed, direct and diffuse solar radiation, and, where missing, outdoor air temperature) are added to the dataset using the Historical Weather API provided by Open-Meteo [29] and by matching the heat pump to its corresponding postal code coordinates. To represent periodic patterns [45], various cyclical variables are generated using sine and cosine functions [9]. Working days, weekdays, season, and year labels are included using integer values (one-hot encoding) [46]. Past load values (lagged values) ranging from 24 hours to 7 days are added. To account for the building's thermal inertia, delayed temperature values are incorporated [45, 47]. Finally, the heat pumps' rated power of the respective time series is included as additional information in the dataset, and added to the feature set.

To simplify model design, no adaptations have been made for forecasting different ASHP load profiles. The final feature set includes only features with Spearman correlation coefficients greater than 0.1 [43] and those that rank in the top 10 in at least 50 percent of the data series according to the Random Forest analysis. Heat pump rated power, as well as the sine and cosine of the time of day, are manually added to the feature set.

The final feature set consists of the following features: average load of the previous day, outdoor temperature, average load at the same time step over the past seven days, lagged temperature (48 h), rolling average temperature (6 h, 12 h, 24 h, 48 h), past load values (24 h, 7 d), relative humidity, sine and cosine of the day, and heat pump size.

4.3 Model Training

4.3.1 XGB: XGB, introduced by [23], is an ensemble learning technique that is based on the *Gradient Boosting Regression Tree* (GBRT) algorithm. In contrast to classical GBTR algorithms, XGB features mathematical and technical improvements, as well as parallel data processing. Due to its high computational efficiency and ability to handle tabular and multivariate data, XGB is one of the most popular methods for load forecasting [18]. The GBRT algorithm sequentially constructs decision trees, splitting the input data at each step based on different features. During prediction, the algorithm searches for the split that represents the best forecast. Each tree corrects the errors of its predecessor, and the final prediction is obtained by summing the weighted predictions of the individual trees. The XGB model is implemented using the open-source Python library *XGB*. Hyperparameters are optimized, while the selection of hyperparameters is based on comparable literature [4, 9, 32]. Hyperparameter tuning is performed using RandomizedSearchCV from the Python library *scikit-learn* [4]. Early stopping is enabled using the root mean squared error (RMSE). The standard error metric, mean squared error (MSE), provided by the Python library is used for model training and to trigger early stopping during validation.

4.3.2 Transformer: As noted in 3, time series forecasting using the Transformer architecture has received attention in the literature. Transformer models process entire data sequences in parallel, in contrast to RNN-based models, which handle sequential data step by step [31]. In particular, their ability to attend to all elements of a sequence enables Transformer models to capture long-term dependencies and complex patterns in sequential data, thereby improving generalization performance [34]. The core principle underlying Transformer models is the attention mechanism, which enables the model to focus on the most relevant parts of a sequence, weigh information by its importance, and learn relationships across long time periods [14]. As a result, Transformers can effectively capture long-range dependencies and leverage information from distant time steps, leading to more robust and flexible predictions.

The model is implemented as a sequence-to-sequence Transformer with an encoder–decoder architecture. The encoder processes load data from the preceding seven days as input to learn temporal dependencies relevant to the target variable, electrical power. The decoder uses auxiliary input features alongside information from the encoder to generate forecasts. Using a seven-day input window, the model predicts the load for the next 24 hours in a single forecasting step (one-step prediction) [18]. Only sequences that are free of temporal discontinuities caused by missing load values are processed. The implementation is based on the *PyTorch* machine learning framework. Prior to training, input features are scaled using a Robust Scaler to mitigate the influence of extreme values. Model performance is evaluated using the Huber loss. The ASHP load time series often exhibit a strong imbalance between very large values (peaks) and much smaller baseline values. The Huber loss is well-suited for this scenario because it combines the advantages of mean absolute error (MAE) and mean squared error (MSE). It penalizes small errors quadratically and large errors linearly, making it less sensitive to outliers while still accounting for deviations across the entire value range [38]. During training, validation, and prediction, only sequences without time gaps are used.

4.4 Benchmark Model

To evaluate the performance of the introduced machine learning models, multiple linear regression and simple daily persistence serve as benchmarks. Forecasting models whose accuracy is lower than the benchmark model are considered to have poor performance [14, 48]. Linear regression models define a direct relationship between the forecast value and the historical values, assuming a linear dependency [16]. A persistence model assumes that the value of the target variable at time step t_i is equal to the value of the target variable at time step t_{i-1}. Persistence methods are suitable when the data exhibit a strong single autocorrelation in lagged values [18].

The linear regression model generates a one-step forecast for the next 24 hours, while the persistence model simply uses the load value from 24 hours earlier. For linear regression, the *LinearRegression* module from the Python library *scikit-learn* is used.

4.5 Metrics

Given the wide range of ASHP load values, commonly used interpretable evaluation metrics are employed to compare results. These include the coefficient of determination (R^2), the normalized RMSE ($nRMSE$), and a peak error metric that sums up not predicted peak loads [49]. In addition, the forecasted load profiles are visually compared. R^2 indicates how superior the model performs compared to a baseline prediction using the mean of the target variable. Higher R^2 values indicate better predictive performance, with $R^2 = 1$ corresponding to a perfect agreement between predictions and observations. Negative R^2 values imply that the model performs worse than the mean-based baseline. However, R^2 should be interpreted with caution, as it may yield deceptively positive values when predictions happen to align closely with the target values by chance or when overfitting causes the model to reproduce the training data well without capturing true underlying patterns [4, 16]. $nRMSE$ is particularly suitable for comparisons at low energy consumption levels, as it is sensitive to errors in small value ranges. For normalization, the value range is used as the scaling factor [41].

4.6 Implementation

Each ASHP time series is split into training, validation, and test sets in a 70:10:20 ratio. Initially, a global XGB model is trained on the entire dataset, meaning that one single model processes multiple load time series simultaneously [10, 14]. In contrast, a local XGB model processes a single time series, so forecasting n load series requires training n separate models. Hyperparameters of the global XGB model are optimized by evaluating 300 random parameter combinations on the test set, with each combination undergoing cross-validation on the validation set. The selection of hyperparameters is based on [9], [4], and [32]. Further trials are performed using a local approach on selected ASHP datasets. The employed hyperparameter sets for each model are provided in Tables via Zenodo https://doi.org/10.5281/zenodo.18369208.

The Transformer model is implemented exclusively as a local model, since a global approach would entail prohibitive computational cost for this load forecasting task. Initially, a simple model architecture with uniform hyperparameters is applied and trained sequentially on each ASHP time series. In subsequent experiments, individual time series are analyzed using optimized hyperparameters.

The machine learning models are trained on an NVIDIA RTX 4000 Ada Generation GPU, while the linear model is trained on the CPU.

5 Forecast assessment

The study compares a globally trained XGB model with locally trained XGB models for selected time series and locally trained Transformer models, using linear regression and persistence as benchmark methods.

5.1 XGB

For the global XGB model with optimized hyperparameters, trained on the entire training set, the distributions of R^2 and $nRMSE$ across all series are approximately unimodal, as shown in Figure 1. Most devices cluster around $R^2 \approx 0.15$, with $nRMSE$ values predominantly between 0.06 and 0.15; only a subset of series with regular, high-amplitude peaks achieves R^2 values above 0.7. In addition, Figure 2 shows the monthly distribution of the R^2 and $nRMSE$ metrics for comparison with the Transformer model. Detailed inspection of individual forecasts demonstrates that the model can successfully capture daily peaks when they occur at nearly fixed times (e.g., around 01:30 in EOH2722), whereas it largely reverts to a smoothed baseline for devices with irregular or user-driven operation (e.g., in KN4), shown in Figure 3.

Local XGB models are trained on three representative series with high, medium, and low predictability to assess whether device-level specialization can substantially improve performance. The results indicate that local models yield only modest gains: for EOH2504 (time series with a consistent pattern), R^2 increases from 0.801 to 0.836 and $nRMSE$ decreases from 0.046 to 0.042, while improvements for the other series remain below four percent. At the same time, the cumulative

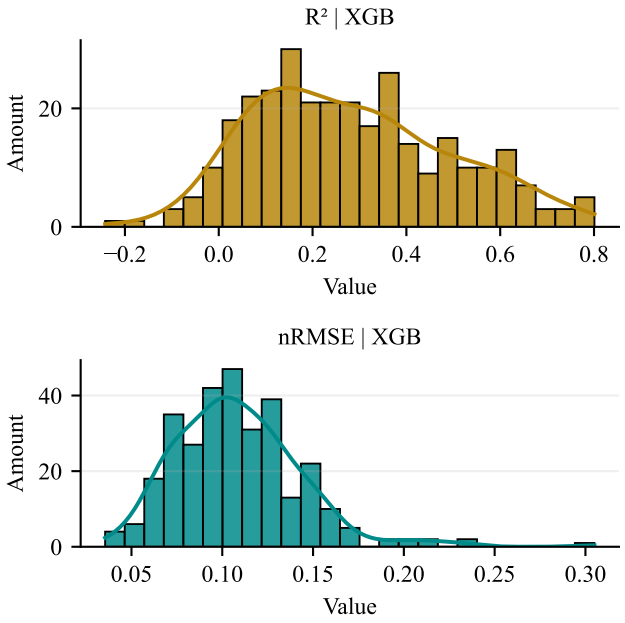

Fig. 1 Comparison of R² and nRMSE for the global XGB Model

optimization time for local models across all devices would significantly exceed that of the global approach. Consequently, the global XGB model appears preferable from a computational efficiency perspective, unless the heterogeneity of the load profiles is exceptionally high.

5.2 Transformer

A simple Transformer architecture with low model complexity is trained sequentially for all series, requiring roughly 22 hours of total training time, which precludes comprehensive hyperparameter tuning across all devices. The resulting R^2 values, shown in Figure 2, display a wide spread: some series achieve monthly R^2 values close to 1, while others fall clearly below the performance of the global XGB model, with overall higher variance in both R^2 and $nRMSE$. Across the device population, some time series exhibit pronounced, quasi-periodic daily peaks, while others are characterized by irregular, noise-dominated patterns with sparse or non-recurrent peaks. When evaluating all models and metrics jointly, this heterogeneity results in substantial variability in forecasting performance, with R^2 values ranging from clearly negative to above 0.8, depending on the underlying load pattern.

To quantify the effect of model complexity, a targeted hyperparameter optimization with *Optuna* is conducted for three selected series. The search explores a broad range of architectural parameters over 75 trials, each trained for up to five epochs with early stopping. The optimized models are consistently more complex than the simple Transformer, yet the performance gains remain marginal. For EOH2504, R^2 improves only from 0.79 to 0.80 and $nRMSE$ from 0.068 to 0.066; for the highly noisy series EOH0793, neither R^2 nor $nRMSE$

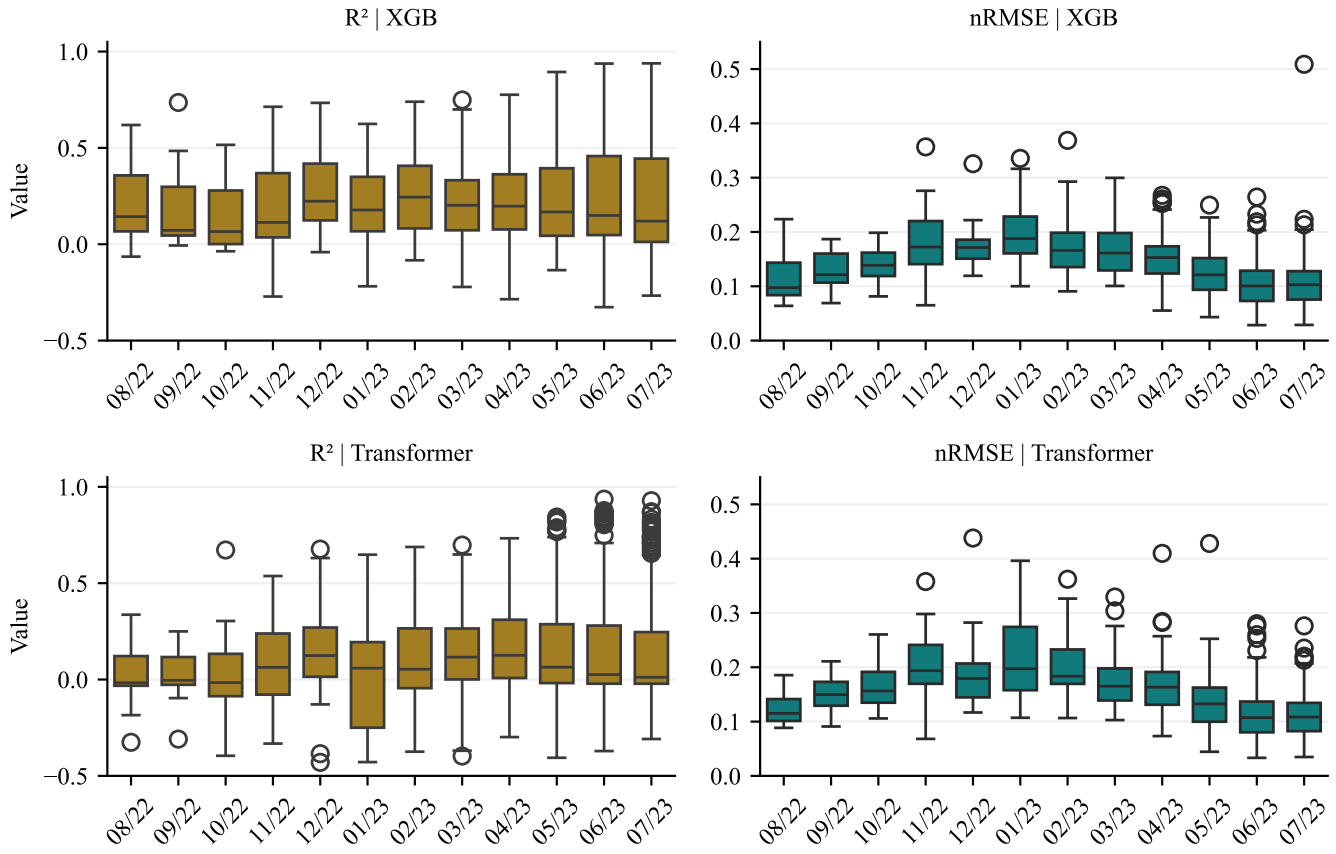

Fig. 2 Monthly comparison of the metric R^2 and $nRMSE$ between the global XGB model (on top) and the Transformer models (on the bottom)

improves substantially. Interestingly, the most complex Transformer configurations are obtained for the most irregular series, yet forecast quality remains low, suggesting that the additional model capacity is used mainly to fit noise rather than to uncover stable patterns. This is further supported by the observation that the loss curves reveal that the best validation performance is already reached after the first epoch, followed by a slight increase in validation loss, indicating a strong tendency to overfit limited and noisy structure.

5.3 Model Comparison

A direct comparison of the machine learning models on two relatively consistent series (EOH1700 and EOH2504) shows that XGB and the Transformer clearly outperform persistence and linear regression, while XGB systematically yields slightly better R^2 and $nRMSE$ than the Transformer. The results are provided in Table 1. For EOH1700, XGB achieves $R^2 = 0.81$ and $nRMSE = 0.064$ compared to $R^2 = 0.80$ and $nRMSE = 0.066$ for the Transformer; for EOH2504, XGB reaches $R^2 = 0.838$ and $nRMSE = 0.041$, while the Transformer achieves $R^2 = 0.775$ and $nRMSE = 0.048$. The peak errors ($Peak_E$) are substantially lower for both machine learning models than for the benchmark methods, especially for series with stable daily patterns.

5.4 Feature Importance

Feature sensitivity analyses confirm that lagged load values are the dominant predictors for both XGB and the Transformer. For EOH2504, removing past load values from the feature set reduces R^2 to approximately -0.03 for both models, whereas removing temperature or cyclic time-of-day features has only

Table 1 Comparison of forecasting methods via different error metrics

Model	R^2	$nRMSE$	$Peak_E$
	EOH1700		
Lin. Regression	0.74	0.075	798.16 kW
Transformer	0.80	0.066	**713.90 kW**
XGB	**0.81**	**0.064**	777.52 kW
Persistenz	0.55	0.099	917.87 kW
	EOH2504		
Lin. Regression	0.408	0.078	358.10 kW
Transformer	0.775	0.048	293.25 kW
XGB	**0.838**	**0.041**	**246.08 kW**
Persistence	0.587	0.078	358.10 kW

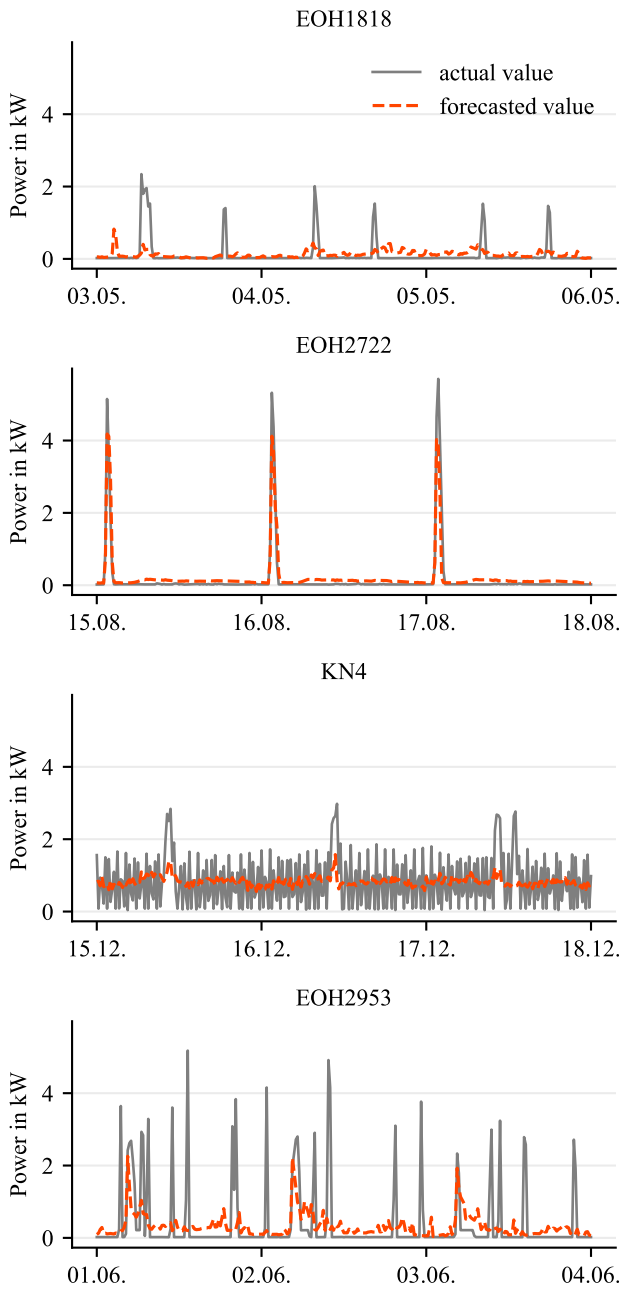

Fig. 3 True and forecasted values for selected time series in summer on top and winter at the bottom

minor effects on R^2. In contrast, models trained only on temperature or only on periodic features perform poorly and fail to capture peaks, supporting the conclusion that these models primarily extrapolate recent load patterns and exploit only limited periodic structure.

5.5 Generalization Analysis

Generalization experiments on an unseen, yet structurally similar series (EOH2722) highlight the dependence of model performance on the regularity of the training data. The results

Table 2 Forecasting unknown time series

Metric	Transf.	XGB	LinReg	Persist.
R^2 Consist.	0.428	**0.434**	0.398	0.235
R^2 Inconsist.	0.038	0.164	0.213	**0.235**
$nRMSE$ Consist.	0.092	**0.091**	0.095	0.107
$nRMSE$ Inconsist.	0.119	**0.111**	0.108	0.107
$Peak_E$ Consist.	14,2k	11,4k	11,6k	**10,4k**
$Peak_E$ Inconsist.	12,2k	14,0k	13,0k	**10,4k**

Unit Peak-Error in MW (k = 1000)

are provided in Table 2. Models trained on a consistent series (EOH2504) achieve $R^2 \approx 0.43$ and $nRMSE \approx 0.09$ on EOH2722 and accurately reproduce regular peaks in both summer and winter, while sporadic peaks remain largely undetected. In contrast, models trained on an irregular series (EOH0793) yield considerably lower R^2 and higher $nRMSE$ on EOH2722, and in some cases are outperformed by the persistence baseline. This is supported by the visualization of a winter and summer week in Figure 4.

5.6 Aggregation Analysis

In an aggregated setting, the average load of the UK time series is used for model training to reduce the impact of data gaps, rather than the aggregated load. The Transformer model slightly outperforms XGB in R^2 and $nRMSE$, while XGB has shorter training times and lower complexity. Linear regression performs worst, especially in summer, underscoring the importance of nonlinear modeling for aggregated yet behavior-influenced loads. Overall, the results are consistent with the well-known aggregation effect: as individual fluctuations average out, the load becomes more predictable, and advanced models can exploit more stable relationships between features and the target.

6 Limitations and Future Work

A central limitation of the study is the restriction to a specific combination of hardware, climate, and usage context, namely residential air-source heat pumps in the UK and southern Germany, which may not be representative for other regions, building typologies, or system configurations such as hybrid or ground-source systems. Moreover, the dataset lacks detailed operational and behavioral information (e.g., thermostat settings, occupancy schedules, or hot-water demands), so the models must infer complex control and user interactions solely from historical power and basic weather data. This missing information contributes to weak performance on irregular devices, whereas it could be made available in energy management systems in the future.

In addition, while the XGB model was extensively optimized and trained on the entire training dataset, the Transformer was only trained and coarsely optimized on one profile at a time due to computational constraints. This implies possible performance gains for the Transformer architecture under

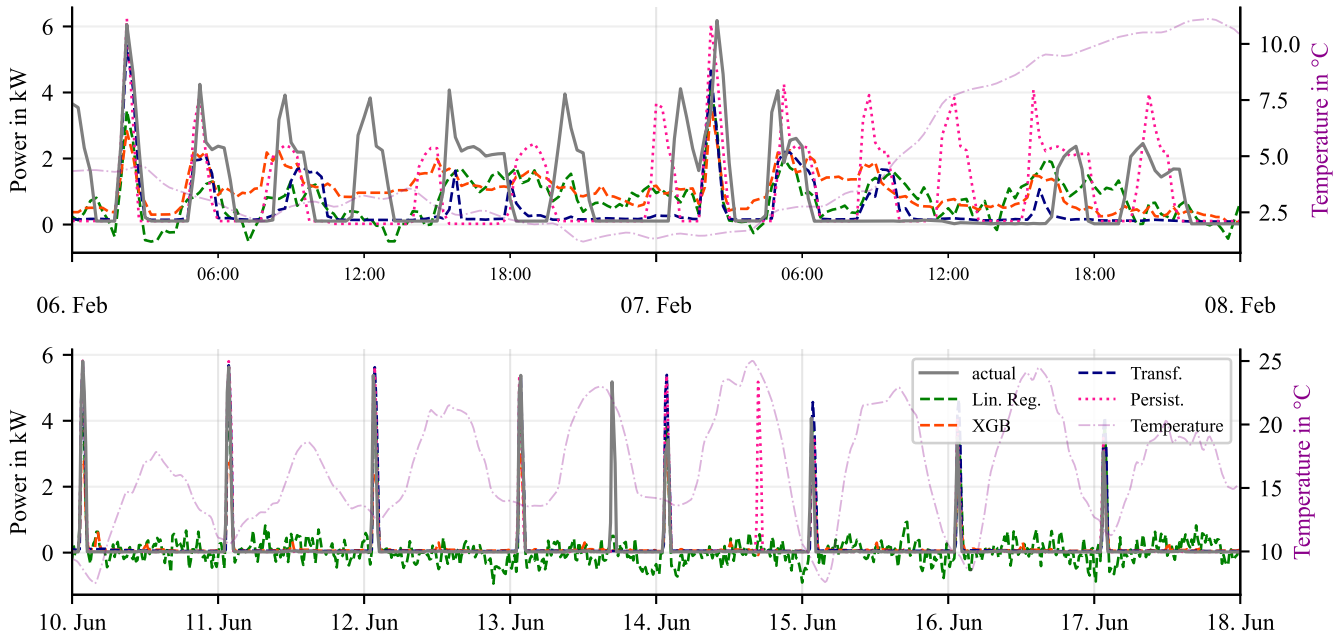

Fig. 4. Comparison of forecasting unknown consistent time series [EOH2722] on consistent trained data [EOH2504]

the same conditions. Finally, the study considers a single-day forecast horizon with fixed 15-minute resolution and does not explore multi-horizon or multi-scale formulations that may better match certain operational use cases.

Future work should therefore extend the analysis in several directions. First, incorporating richer contextual information, such as indoor temperatures, occupancy proxies, tariff signals, or hot water demand indicators, may enable models to separate user behaviour from building and system dynamics and improve forecasts, particularly for irregular devices. Second, hierarchical and transfer learning approaches that leverage similarities across devices (e.g., by clustering heat pumps by usage archetypes and training shared models with device-specific adaptations) could mitigate data sparsity and improve generalization to new installations. Finally, future work should explicitly link forecast performance to downstream control performance in realistic energy management simulations, thereby quantifying the operational value of improved forecasts at both device and aggregated levels.

7 Conclusion

The empirical analysis demonstrates that device-level day-ahead forecasting of residential ASHP load at 15-minute resolution is fundamentally constrained by the underlying heterogeneity of load profiles and user-driven operation. While both XGB and Transformer models outperform linear regression and persistence for devices with stable, quasi-periodic load patterns, XGB consistently achieves slightly higher R^2 and lower $nRMSE$ at substantially lower computational cost, making it a practical choice for large device populations. For irregular, behavior-dominated time series, however, all considered methods deliver modest accuracy and often fail to capture sporadic peaks, indicating that model class alone cannot compensate for low intrinsic predictability.

Feature importance and ablation analyses show that recent load history is the key driver of forecast accuracy, while exogenous factors such as temperature and cyclic calendar variables mainly provide incremental refinements and cannot compensate for missing behavioural information. Generalization experiments further indicate that models tend to inherit the structural properties of their training series, which limits the transferability of models trained on noisy loads and underscores the importance of carefully selecting training data for deployment in new devices. From a system perspective, these findings suggest that while device-level forecasts can support Prosumer energy management for a subset of heat pumps under regular operation, aggregated forecasting at the building, cluster, or community scale provides more reliable and robust inputs for assessing grid impacts and exploiting flexibility in highly electrified residential sectors.

8 Acknowledgements

This research was funded by "zukunft.niedersachsen", the joint science funding program of the Lower Saxony Ministry of Science and Culture and the Volkswagen Foundation, under project grant number ZN4462. The content of this paper reflects the authors' views and responsibilities solely.

9 References

[1] Appunn, K., Eriksen, F. & Wettengel, J. Germany's greenhouse gas emissions and energy transition targets. , https://www.cleanenergywire.org/factsheets/germanys-greenhouse-gas-emissions-and-climate-targets

[2] Wirth, H. Recent Facts about Photovoltaics in Germany. , https://www.pv-fakten.de/

[3] Haendel, M., Hug, G. & Klobasa, M. Effects of heat pump scheduling on low–voltage grids using a receding horizon control strategy. *IET Smart Grid*. **6**, 432-445 (2023)

[4] Song, Y., Peskova, M., Rolando, D., Zucker, G. & Madani, H. Estimating electric power consumption of in-situ residential heat pump systems: A data-driven approach. *Applied Energy*. **352** pp. 121971 (2023,12)

[5] Xie, Y., Hu, P., Zhu, N., Lei, F., Xing, L., Xu, L. & Sun, Q. A hybrid short-term load forecasting model and its application in ground source heat pump with cooling storage system. *Renewable Energy*. **161** pp. 1244-1259 (2020,12), https://www.sciencedirect.com/science/article/pii/S0960148120312180

[6] Xu, C., Chen, H., Xun, W., Zhou, Z., Liu, T., Zeng, Y. & Ahmad, T. Modal decomposition based ensemble learning for ground source heat pump systems load forecasting. *Energy And Buildings*. **194** pp. 62-74 (2019,7), https://www.sciencedirect.com/science/article/pii/S0378778818335631

[7] Esen, H., Inalli, M., Sengur, A. & Esen, M. Forecasting of a ground-coupled heat pump performance using neural networks with statistical data weighting pre-processing. *International Journal Of Thermal Sciences*. **47**, 431-441 (2008,4), https://www.sciencedirect.com/science/article/pii/S1290072907000762

[8] Schmitz, S., Brucke, K., Kasturi, P., Ansari, E. & Klement, P. Forecast-based and data-driven reinforcement learning for residential heat pump operation. *Applied Energy*. **371** pp. 123688 (2024,10), https://www.sciencedirect.com/science/article/pii/S0306261924010717

[9] Semmelmann, L., Hertel, M., Kircher, K., Mikut, R., Hagenmeyer, V. & Weinhardt, C. The impact of heat pumps on day-ahead energy community load forecasting. *Applied Energy*. **368** pp. 123364 (2024)

[10] Hertel, M., Beichter, M., Heidrich, B., Neumann, O., Schäfer, B., Mikut, R. & Hagenmeyer, V. Transformer training strategies for forecasting multiple load time series. *Energy Informatics*. **6**, 20 (2023,10), https://doi.org/10.1186/s42162-023-00278-z

[11] Wei, Z., Zhang, T., Yue, B., Ding, Y., Xiao, R., Wang, R. & Zhai, X. Prediction of residential district heating load based on machine learning: A case study. *Energy*. **231** pp. 120950 (2021)

[12] Ji, Y., Buechler, E. & Rajagopal, R. Data-Driven Load Modeling and Forecasting of Residential Appliances. *IEEE Transactions On Smart Grid*. **11**, 2652-2661 (2020)

[13] Razghandi, M. & Turgut, D. Residential Appliance-Level Load Forecasting with Deep Learning. *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*. pp. 1-6 (2020,12), https://ieeexplore.ieee.org/document/9348197, ISSN: 2576-6813

[14] Haben, S., Voss, M. & Holderbaum, W. Core Concepts and Methods in Load Forecasting: With Applications in Distribution Networks. (Springer International Publishing,2023)

[15] Roy, S., Samui, P., Nagtode, I., Jain, H., Shivaramakrishnan, V. & Mohammadi-ivatloo, B. Forecasting heating and cooling loads of buildings: a comparative performance analysis. *Journal Of Ambient Intelligence And Humanized Computing*. **11**, 1253-1264 (2020,3), https://doi.org/10.1007/s12652-019-01317-y

[16] Paravantis, J., Malefaki, S., Nikolakopoulos, P., Romeos, A., Giannadakis, A., Giannakopoulos, E., Mihalakakou, G. & Souliotis, M. Statistical and machine learning approaches for energy efficient buildings. *Energy And Buildings*. **330** pp. 115309 (2025,3), https://www.sciencedirect.com/science/article/pii/S0378778825000398

[17] Ma, P., Cui, S., Chen, M., Zhou, S. & Wang, K. Review of Family-Level Short-Term Load Forecasting and Its Application in Household Energy Management System. *Energies*. **16**, 5809 (2023,1), https://www.mdpi.com/1996-1073/16/15/5809, Number: 15 Publisher: Multidisciplinary Digital Publishing Institute

[18] Hou, H., Liu, C., Wang, Q., Wu, X., Tang, J., Shi, Y. & Xie, C. Review of load forecasting based on artificial intelligence methodologies, models, and challenges. *Electric Power Systems Research*. **210** pp. 108067 (2022,9), https://www.sciencedirect.com/science/article/pii/S0378779622002917

[19] Runge, J. & Zmeureanu, R. A Review of Deep Learning Techniques for Forecasting Energy Use in Buildings. *Energies*. **14**, 608 (2021,1), https://www.mdpi.com/1996-1073/14/3/608, Number: 3 Publisher: Multidisciplinary Digital Publishing Institute

[20] Sun, Y., Haghighat, F. & Fung, B. A review of the-state-of-the-art in data-driven approaches for building energy prediction. *Energy And Buildings*. **221** pp. 110022 (2020,8)

[21] Eren, Y. & Küçükdemiral, İ. A comprehensive review on deep learning approaches for short-term load forecasting. *Renewable And Sustainable Energy Reviews*. **189** pp. 114031 (2024,1), https://www.sciencedirect.com/science/article/pii/S1364032123008894

[22] Hagan, M. & Behr, S. The Time Series Approach to Short Term Load Forecasting. *IEEE Transactions On Power Systems*. **2**, 785-791 (1987), http://ieeexplore.ieee.org/document/4335210/

[23] Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings Of The 22nd ACM SIGKDD International Conference On Knowledge Discovery And Data Mining*. pp. 785-794 (2016,8), arXiv:1603.02754 [cs]

[24] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. & Polosukhin, I. Attention Is All You Need. *NIPS'17: Proceedings Of The 31st International Conference On Neural Information Processing Systems*. (2017,12), arXiv:1706.03762 [cs]

[25] Chollet, F. Deep Learning with Python. (Manning Publications Co.,2021)

[26] Energy Systems Catapult Electrification of Heat Demonstration Project: Cleansed 2-Minute Interval Heat Pump Performance Data. (UK Data Archive,2020), https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=9050

[27] Energy Systems Catapult Electrification of Heat Demonstration Project: Heat Pump Performance Cleansed Data, 2020-2023. (UK Data Service,2024)

[28] CoSSMic Open Power System Data. (2020,4), https://data.open-power-system-data.org/household_data/2020-04-15/

[29] Zippenfenig, P. Open-Meteo.com Weather API. (Zenodo,2023), https://doi.org/10.5281/ZENODO.7970649

[30] Josep Ferrer Wie Transformatoren funktionieren: Eine detaillierte Erkundung der Transformatorarchitektur. *Datacamp*. (2024,9), https://www.datacamp.com/de/tutorial/how-transformers-work

[31] Chan, J. & Yeo, C. A Transformer based approach to electricity load forecasting. *The Electricity Journal*. **37**, 107370 (2024,3), https://www.sciencedirect.com/science/article/pii/S1040619024000058

[32] Wang, Z., Hong, T. & Piette, M. Building thermal load prediction through shallow machine learning and deep learning. *Applied Energy*. **263** pp. 114683 (2020,4)

[33] Wei, D. Essential Python for Machine Learning: XGBoost. *Medium*. (2024,1), https://medium.com/@weidagang/essential-python-for-machine-learning-xgboost-4b662cf19fcd

[34] Wang, C., Wang, Y., Ding, Z., Zheng, T., Hu, J. & Zhang, K. A Transformer-Based Method of Multienergy Load Forecasting in Integrated Energy System. *IEEE Transactions On Smart Grid*. **13**, 2703-2714 (2022,7), Conference Name: IEEE Transactions on Smart Grid

[35] Bergmann, D. & Stryker, C. Was ist PyTorch?. (2025), https://www.ibm.com/de-de/think/topics/pytorch

[36] PyTorch PyTorch Forecasting Documentation. *PyTorch-Forecasting*. (2025), https://pytorch-forecasting.readthedocs.io/en/stable/#

[37] Yadav, A. Time Series Forecasting with PyTorch. *Medium*. (2024,10), https://medium.com/we-talk-data/time-series-forecasting-with-pytorch-c18fc512daf4

[38] Chen, B. Understanding Huber Loss function: Insights from Applications. *Medium*. (2024,6), https://medium.com/@devcharlie2698619/understanding-huber-loss-function-insights-from-applications-5c1c5145d2c4

[39] Kamel, E., Sheikh, S., & Huang, X. Data-driven predictive models for residential building energy use based on the segregation of heating and cooling days. *Energy, Volume 206*. (2020), https://doi.org/10.1016/j.energy.2020.118045

[40] Cholewa, T., Siuta-Olcha, A., Smolarz, A., Muryjas, P., Wolszczak, P., Guz, Ł. & Balaras, C. On the short term forecasting of heat power for heating of building. *Journal Of Cleaner Production*. **307** pp. 127232 (2021,7)

[41] Lusis, P., Khalilpour, K., Andrew, L. & Liebman, A. Short-term residential load forecasting: Impact of calendar effects and forecast granularity. *Applied Energy*. **205** pp. 654-669 (2017,11), https://www.sciencedirect.com/science/article/pii/S0306261917309881

[42] Koprinska, I., Rana, M. & Agelidis, V. Correlation and instance based feature selection for electricity load forecasting. *Knowledge-Based Systems*. **82** pp. 29-40 (2015,7), https://www.sciencedirect.com/science/article/pii/S0950705115000714

[43] Zhang, L., Wen, J., Li, Y., Chen, J., Ye, Y., Fu, Y. & Livingood, W. A review of machine learning in building load prediction. *Applied Energy*. **285** pp. 116452 (2021,3), https://www.sciencedirect.com/science/article/pii/S0306261921000209

[44] Iranzad, R. & Liu, X. A review of random forest-based feature selection methods for data science education and applications. *International Journal Of Data Science And Analytics*. (2024,2), https://doi.org/10.1007/s41060-024-00509-w

[45] Mayrink, V. & Hippert, H. A hybrid method using Exponential Smoothing and Gradient Boosting for electrical short-term load forecasting. *2016 IEEE Latin American Conference On Computational Intelligence (LA-CCI)*. pp. 1-6 (2016,11), https://ieeexplore.ieee.org/document/7885697

[46] Eseye, A. & Lehtonen, M. Short-Term Forecasting of Heat Demand of Buildings for Efficient and Optimal Energy Management Based on Integrated Machine Learning Models. *IEEE Transactions On Industrial Informatics*. **16**, 7743-7755 (2020,12), https://ieeexplore.ieee.org/document/8990012

[47] Sun, Y., Haghighat, F. & Fung, B. A review of the-state-of-the-art in data-driven approaches for building energy prediction. *Energy And Buildings*. **221** pp. 110022 (2020,8)

[48] Koukaras, P., Bezas, N., Gkaidatzis, P., Ioannidis, D., Tzovaras, D. & Tjortjis, C. Introducing a novel approach in one-step ahead energy load forecasting. *Sustainable Computing: Informatics And Systems*. **32** pp. 100616 (2021,12), https://www.sciencedirect.com/science/article/pii/S2210537921001049

[49] Melillo, A., Meyer, M., Hendry, R. & Schuetz, P. Prediction of heat pump demand profiles with few Dependencies: A combination of statistical and physical modelling. *Energy And Buildings*. **338** pp. 115712 (2025,7), https://www.sciencedirect.com/science/article/pii/S0378778825004426