Data-Driven Approach to Predict the Consumption of Electrical Energy in Households Using Features from Non-Electric Data

Fábio Sayeg
Instituto Superior Técnico - IST, U. Lisbon
Lisbon, Portugal
fabio.sayeg@tecnico.ulisboa.pt

João Góis
ITI / LARSyS and IST, U. Lisbon
Lisbon, Portugal
joao.gois@tecnico.ulisboa.pt

Lucas Pereira
ITI / LARSyS and IST, U. Lisbon
Lisbon, Portugal
lucas.pereira@tecnico.ulisboa.pt

***Abstract*—This paper addresses a research gap in predicting residential energy consumption by proposing a data-driven approach utilizing features from non-electric data for training the machine-learning models. Specifically, forecasting models are trained to predict the aggregated household demand for one day, seven days, and one month. Comparisons are made to forecasting models trained on historical consumption data. By employing machine learning algorithms and exploring two distinct approaches - utilizing dwelling and occupants' data with seasonal factors (non-electric data) and historical time-series consumption data - the study provides valuable insights into energy consumption prediction based on household characteristics. Results using data from 20 households in the UK indicate that while utilizing historical consumption data yields superior performance, the proposed approach remains a viable alternative in cases where historical consumption time series are unavailable, demonstrating promising results for forecasting household energy demand.**

***Index Terms*—Household Demand; Forecasting; Machine-Learning; Non-Electric Data; Historical Consumption**

I. Introduction

The surge in residential energy consumption within the European Union (EU) over recent years underscores the need for innovative strategies to manage this escalating demand. Factors such as heightened expectations for comfort, larger living spaces, and the proliferation of new appliances have all contributed to this uptick in usage. Consequently, the imperative to transition towards cleaner energy sources has gained traction to curb greenhouse gas emissions [1].

Renewable energies, particularly solar power, have emerged as viable alternatives, supported by their increasing accessibility and affordability [2]. However, effectively integrating Renewable Energy Systems (RES) necessitates a nuanced understanding of residential energy consumption dynamics, a task often complicated by its intricate nature [3].

Today, the widespread adoption of smart meters provides a solution for monitoring household energy consumption [4]. In the United States, for instance, over 111 million smart meters were in use in 2021, with about 88% of them used in residences [5]. These devices offer highly detailed data, enabling consumers to perform comprehensive energy analysis [4], [6]. Nevertheless, utility companies must use smart meters with caution due to the sensitive nature of the data regarding users' lifestyles and habits [7].

Consequently, despite the widespread adoption of smart meters, historical data is often unavailable. Ultimately, this challenges developing algorithms to estimate energy consumption using traditional machine-learning approaches that most often depend on the availability of vast amounts of historical time series analysis consumption data [8].

One approach to circumvent these challenges involves broadening the scope of analysis to encompass non-electric data sources. Variables such as household demographics, occupancy patterns, appliance usage, building age, geographical location, and weather conditions can provide valuable energy forecasting insights [9]. Still, despite the potential benefits of such an approach, the core of the research is devoted to forecasting energy demand with historical consumption data.

This work contributes to mitigating this important research gap by proposing a data-driven approach to predict residential energy consumption using features from non-electric data. To state more precisely, forecasts will be produced considering three distinct periods: one day, seven days, and one month. The results of the proposed approach are also compared with traditional forecasting algorithms that rely on historical consumption data.

This paper is organized as follows. A brief overview of related works is presented in Section II. The dataset, forecasting methods, and evaluation strategies are described in Section III. The forecasting results are presented and discussed in Section IV. The main conclusions, limitations and future work directions are presented in Section V.

II. Related Works

The research community has extensively investigated the challenge of predicting building energy consumption using historical consumption data. In their comprehensive review, Amasyali et al. [10] offers a broad overview of the literature employing machine learning techniques for this purpose. They categorize each study based on the machine learning algorithms employed, the building types under study, temporal granularity, predicted energy consumption types, prediction purposes, dataset types (real or simulated), considered features, dataset sizes or recording periods, and performance metrics.

Conversely, the exploration of demand forecasting using non-electric features remains relatively limited. Some works resort to clustering techniques to group consumers into different profiles that present a similar behavior or are part of a socio-demographic group [11]–[14]. Others focus on analyzing the impact of physical characteristics of the buildings to have a quantitative view of the problem based on these kinds of features of the dwellings [15], [16]. In this context, Santin [9] predicted the energy used for heating based on household and building and occupant characteristics, lifestyle, and the number of electronics. This prediction analyzed how each of the characteristics influenced the energy consumption of the dwelling and presented a numerical analysis of this impact. In another work [17], the authors concluded that non-electric data features affect the performance of Non-Intrusive Load Monitoring (NILM) algorithms. Although not focused on load demand forecasting, this study shows that non-electric data impacts energy consumption patterns, potentially influencing forecasting accuracy.

III. Materials and Methods

This section presents the dataset and the employed regression algorithms - including the optimized hyper-parameters. The training and testing procedures are also presented.

A. Dataset

To proceed with this work, an essential requirement was to identify a publicly available household energy consumption dataset that included household features. To this end, the REFIT dataset [18] was selected. This dataset contains approximately two years of consumption data and a detailed description of the households, including occupancy, the construction year, size, and the total number of appliances. Table I lists the non-electric features considered in this work.

TABLE I: List of non-electric features used in this work.

Feature	Classification	Feature	Classification
Month	Seasonal	Television_qtt	House
Year	Seasonal	Electric Heater_qtt	House
Weekday	Seasonal	Hi-Fi_qtt	House
Season	Seasonal	Fridge-Freezer_qtt	House
Occupancy	House	Microwave_qtt	House
Construction_Year	House	Kettle_qtt	House
Appliances	House	Toaster_qtt	House
Size	House	Misc Kitchen_qtt	House
Fridge_qtt	House	Tumble Dryer_qtt	House
Freezer_qtt	House	Router_qtt	House
Washer Dryer_qtt	House	Games Console_qtt	House
Washing Machine_qtt	House	Misc_qtt	House
Dishwasher_qtt	House	Cooker Hood_qtt	House
Computer_qtt	House	Unidentified_qtt	House

The REFIT dataset contains consumption records roughly every eight seconds. However, for this work, the resolution was set to one sample per hour by taking the average of the measurements within each hour.

B. Forecasting Algorithms

Various machine learning algorithms were tested, with hyper-parameter tuning conducted via grid search. This technique tests every possible combination of pre-defined values of variables together and evaluates the results in each possible scenario [19]. The forecast time granularity is one day, seven days, and one month.

1) Forecasting with Non-Electric Data: For forecasting with non-electric data, the following six algorithms were considered: K-Nearest Neighbors (KNN), Regression Trees, Random Forests, Regression Trees with Gradient Boosting, Multi-Layer Perceptron (MLP), and Support Vector Regression (SVR). The optimized hyper-parameters are presented in Table II, with the best value for each hyper-parameter highlighted in boldface and underlined.

TABLE II: Hyper-parameters for forecasting algorithms using non-electric data.

Algorithm	Hyper-parameter	Tested Values
kNN	N.º Neighbours	**<u>3</u>** to 19 in steps of 2
	Distance	Euclidean, Manhattan, **<u>Chebyshev</u>**
	Scaling	None, **<u>Min-Max</u>**, Standardization
Regression Tree	Max. Depth	**<u>2</u>**, 5 to 25 in steps of 5
	Min. Impurity	0.01, **<u>0.005</u>**, 0.0025, 0.001, 0.0005
Random Forest	Max. Depth	5, **<u>10</u>**, 25, unlimited
	N.º Estimators	5,**<u>10</u>**,25,50,75,100,22,300,400
	Max.Features	0.3, 0.5, 0.7, **<u>1</u>**
Gradient Boosting	Max. Depth	5, **<u>10</u>**, 25
	N.º Estimators	**<u>50</u>**, 100, 200, 500
	Learning Rate	**<u>0.01</u>**, 0.1, 0.5, 0.9
MLP	Activation Function	ReLu, **<u>Logistic</u>**
	Initial Learning Rate	0.01, 0.1, **<u>0.5</u>**, 0.9
	Max Iterations	**<u>100</u>**, 150, 250, 500, 1000
SVR	Kernel	Linear, **<u>Polynomial</u>**, radial basis function, sigmoid
	C-Value	**<u>0.1</u>**, 1, 10
	Gamma	0,1, **<u>1</u>**, auto, scale

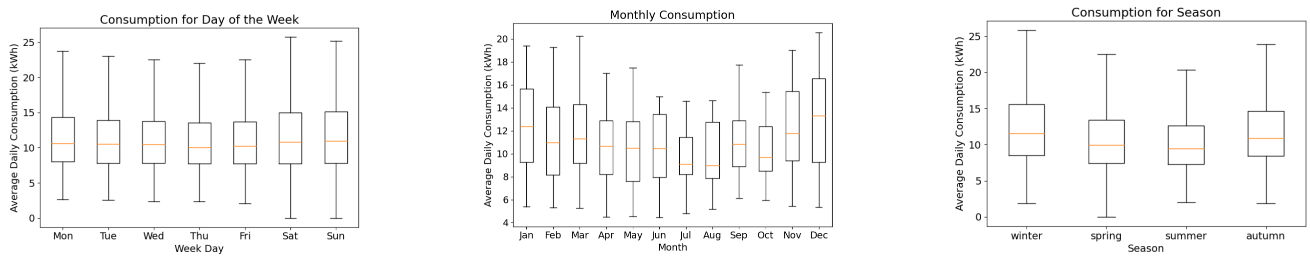
2) Forecasting with Historical Consumption: For the forecasting using time series data, the MLP, the Long-Short Term Memory (LSTM), the Auto-Regressive Integrated Moving Average (ARIMA), and the Seasonal ARIMA (SARIMA) were considered. The fined-tuned hyper-parameters are displayed in Table III.

TABLE III: Hyper-parameters for forecasting algorithms using historical electricity consumption data.

Algorithm	Hyper-parameter	Tested Values
MLP / LSTM	Historical Window	3, 10, 15, 45, 60
	Learning Rate	0.001. 0.01, 0.1
	Dropout rate	0.2, 0.5. 0.8
	Weight Decay	0.001,0.0001, 0

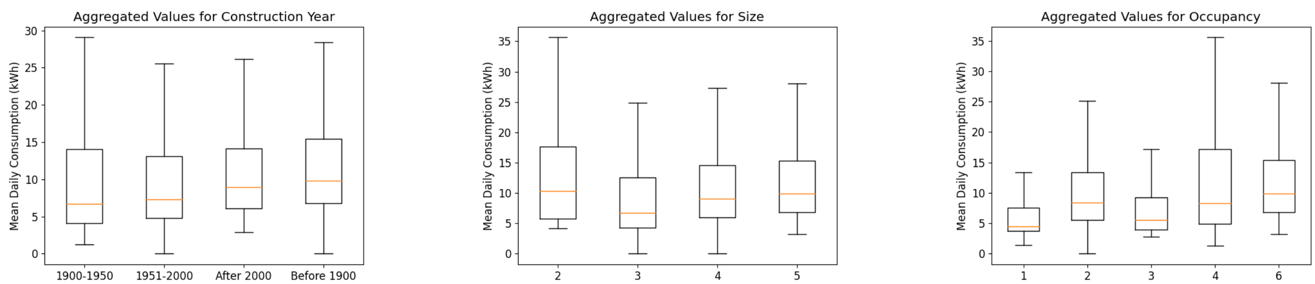
C. Training and Testing

Due to the different nature of the two regression problems, it was necessary to establish two different training and testing setups.



(a) **(b)** **(c)**

Fig. 1: Analysis of seasonal patterns: (a) Weekday patterns, (b) Monthly patterns, (c) Seasons patterns



(a) **(b)** **(c)**

Fig. 2: Analysis of occupational patterns: (a) Construction year, (b) Number of bedrooms, (c) Number of occupants.

1) Non-Electric Data: K-fold cross-validation: To train and test the algorithms that will use non-electric data, it was decided to use k-fold cross-validation. In this process, the training data is divided into K approximately equal parts. The model is then trained on K-1 parts while evaluating the test error using the trained model on the remaining K-th part. This procedure is repeated K times, where each subset of data serves as the test set once [20], [21].

Considering the objective of this work, the idea is that the folds used in the k-folds cross-validation will be the 20 houses of the dataset (hence, k will be equal to 20). The objective here is to reduce the algorithm's bias by not training it using records of the same house of the test. That strategy will also present evaluation test results for each house, enabling individual analysis for each residency of the dataset.

2) Historical Consumption: Train-Test Split: For the algorithms that used historical data for training, it was decided to use a train-test split evaluation strategy, where 70% of the time series is used for training, 20% for validation, and 10% for testing. This methodology was employed individually in each house.

For the MLP and LSTM algorithm, the training, validation, and test sets were further divided into windows with the length of the input samples (historical window size) and outputs (forecasting horizon, which, in this case, is set to one sample). This technique is known as a rolling-origin evaluation [22]. As for the ARIMA and SARIMA, which do not require window generation for prediction, they were trained, validated, and tested using the entire time series for each house, divided into the training, validation, and test sets.

3) Performance Metric: The metric chosen to evaluate the results was the Mean Absolute Error (MAE), as this is a ubiquitous measure to assess the accuracy in regressive models [23]. Besides, since this metric keeps the same scale as the records, it is possible to analyze the magnitude of the error quickly.

IV. Results and Discussion

A. Analysis of Seasonal and Occupancy Patterns

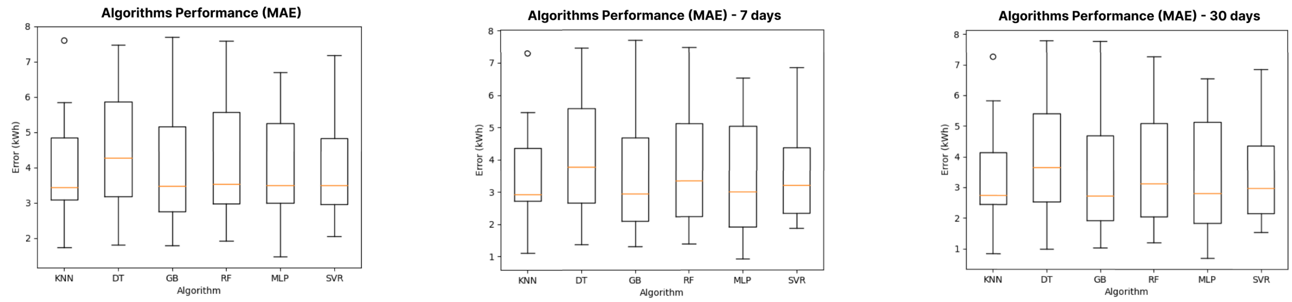
Before delving into machine learning algorithms, it's valuable to analyze potential patterns related to the target variable, including months, days of the week, hours of the day, and seasonal variations.

An initial examination of these patterns reveals several insights. Firstly, as depicted in Figure 1(a), the impact of the day of the week on energy consumption appears slight, with a minor increase during weekends and a decrease during weekdays, particularly at the beginning of the week. This suggests that while the day of the week may influence forecasts somewhat, it may not be among the most significant features.

Moreover, seasonal variations, as illustrated in Figure 1(b) and Figure 1(c), demonstrate that colder months, such as winter and autumn, exhibit higher consumption rates, likely due to increased heating equipment usage.

Other relevant aspects include building and dwelling features, such as the year of construction, number of bedrooms, and number of occupants. The relationships between these features and total energy consumption are depicted in Figures 2(a) to 2(c).

However, deriving clear patterns directly from these features alone is challenging, as the most significant determinants of



(a) **(b)** **(c)**

Fig. 3: Mean Absolute Error for Non-Electric Data: (a) 1-day average, (b) 7-day average, (c) 30-day average.

energy consumption likely stem from occupant habits rather than dwelling features. Nonetheless, the graph in Figure 2(a) suggests a potential tendency for older buildings to require slightly more energy. Interestingly, the graph in Figure 2(b) reveals that occupants of houses with two bedrooms may tend to consume more energy than those in larger houses. This unexpected pattern might be attributed to high variance within this category, indicating specific houses with unusually high consumption.

B. Predictions Using Occupants, Dwelling, and Seasonal Data

Boxplots representing error values at granularities of 1 day, seven days, and 30 days are depicted in Figure 3(a), Figure 3(b), and Figure 3(c) respectively.

From the analysis, it's evident that KNN exhibited favorable performance regarding mean error and variance across all granularities, while Decision Trees performed relatively poorly, displaying the highest average error and variance. Random Forests and Decision Trees with Gradient Boosting showcased similar error profiles with moderately high average errors and variance. The Multi-Layer Perceptron demonstrated average error with high variance, albeit achieving good results for some households. Support Vector Regression displayed average error with slightly lower variance.

Examining the best-performing algorithm in each household provided insights into the impact of reducing time granularity. Graph 4 illustrates that, for most cases, there was a notable reduction in error when transitioning from one day to seven days and from seven days to 30 days granularities. While some households, like 4, 12, and 18, also experienced positive impacts from granularity reduction, it wasn't as significant as others. Houses 14 and 20 exhibited different patterns, suggesting that algorithms perform better at predicting consumption patterns rather than exact consumption on specific dates.

C. Predictions Using Historical Consumption Data

The forecast results using historical data for the different periods can be seen in the boxplots in Figure 5(a), Figure 5(b), and Figure 5(c).

As depicted in Figure 5(a), the Multi-layer Perceptron (MLP) emerged as the top-performing algorithm for one-day consumption forecasts, boasting the lowest average absolute error and demonstrating a stable behavior across all dataset houses. Conversely, the LSTM showed the poorest performance for one-day predictions, characterized by high variance, elevated average percentage error, and notable negative outliers in certain predictions. Both ARIMA and SARIMA showcased average performance, featuring moderate absolute and percentage errors along with typical variance. However, they were prone to numerous outliers, indicating suboptimal performance for certain dataset houses.

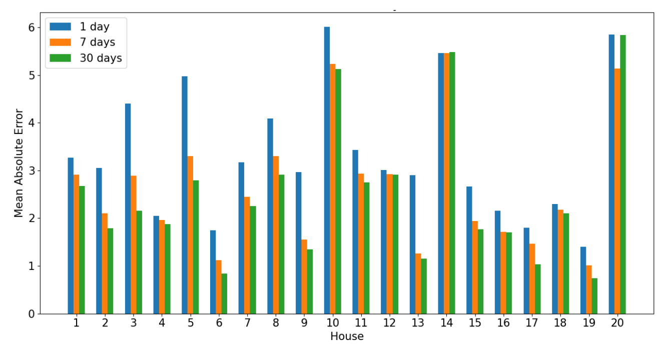
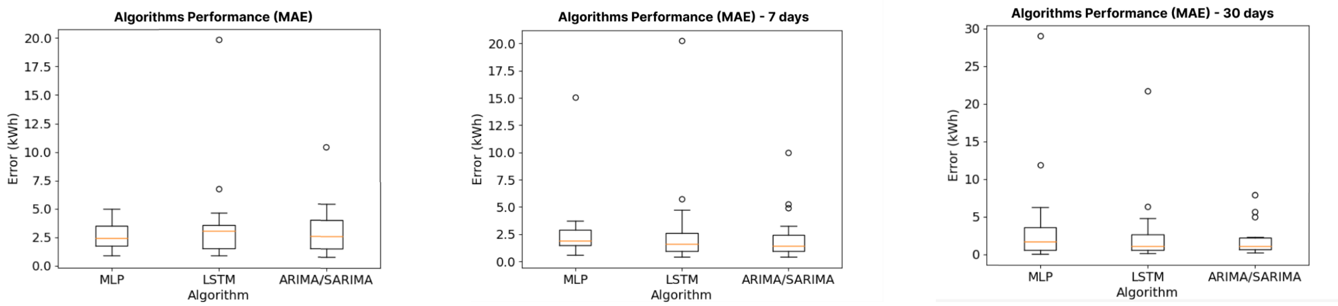


Fig. 4: Non-Electric Data: Mean Absolute Error for three different granularities for all houses (best algorithm).

Similar patterns were observed for seven-day predictions, as illustrated by the boxplots in Figure 5(b). Despite substantial variance, the MLP demonstrated favorable performance, while the LSTM showed slight improvement, particularly in average percentage error. ARIMA and SARIMA exhibited enhanced performance compared to one-day predictions, manifesting lower MAE on average and reduced variance. However, these algorithms still faced challenges with many negative outliers.

Turning to 30-day forecasts, depicted in Figure 5(c), the MLP displayed the poorest performance, with the highest average error and variance. The LSTM exhibited slightly lower error and variance compared to the MLP. Conversely, ARIMA and SARIMA again outperformed other algorithms, showcasing low error rates and the lowest variance.

For a detailed examination of results across individual houses, refer to Figure 6. Generally, there was a tendency



(a) (b) (c)

Fig. 5: Mean Absolute Error with Historical Consumption Data: (a) 1-day average, (b) 7-day average, (c) 30-day average.

for error reduction with decreasing granularity, as seen in the bar plots. However, House 15 displayed a deviation from this trend, suggesting that historical forecasts for longer intervals yielded unfavorable results for this particular household. Additionally, House 20 emerged as a clear outlier, exhibiting consistently poor results across all granularities compared to other dataset houses.

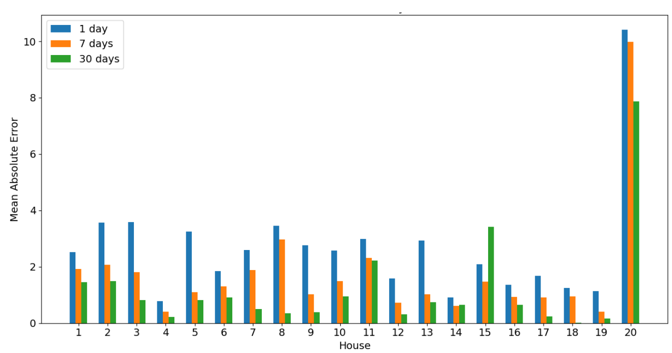


Fig. 6: Historical Consumption Data: Mean Absolute Error for three different granularities for all houses (best algorithm).

D. Discussion

The comparison between predictions based on historical data and household characteristics reveals exciting findings. Neural Networks outperformed other algorithms for one-day energy consumption forecasts when applied to historical time series data, as depicted in Figure 7(a). Similarly, algorithms utilizing historical data generally showed superior performance for seven-day averages (Figure 7(b)). In contrast, for 30-day averages, Neural Networks were less effective than the ARIMA/SARIMA alternative. Interestingly, the disparities between the two forecasting approaches were less pronounced than initially expected, especially for 30-day forecasts (Figure 7(c)).

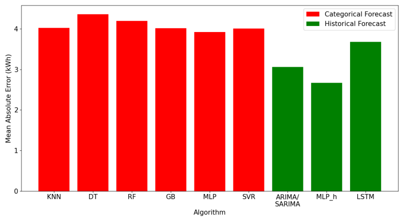
Moreover, the superior performance observed in algorithms utilizing historical data may also be attributed to the training procedure. Training and testing are conducted using data specific to each household in these algorithms. In contrast, training and testing involve entirely different households when forecasting using only non-electric features. Despite this, the differences between the two methods remain relatively modest.

To conclude, while algorithms relying solely on non-electric features may exhibit higher errors than alternatives requiring historical data, leveraging household features for demand prediction can still provide valuable insights and practical utility, particularly in scenarios where historical data is scarce or inaccessible. One particular use case is the development of load models for grid simulation and planning since traditional tools often rely on simplistic assumptions or aggregated data, which may not capture the diversity and complexity of household consumption patterns. By incorporating detailed household feature data into the load modeling task, researchers can create more realistic and granular representations of consumer behavior, leading to more accurate system simulations and predictions.

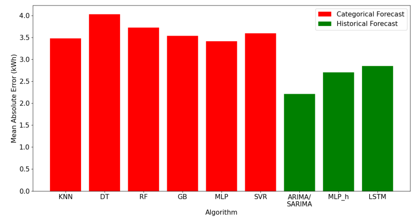
V. Conclusion

The paper successfully addresses the prediction of electrical energy consumption using machine learning algorithms, exploring two distinct approaches: utilizing dwelling and occupants' data with seasonal factors (i.e., non-electric data) and employing historical time-series consumption data. It fills a research gap by delving into energy consumption prediction based on household data, offering valuable insights for future studies on an area that remains relatively unexplored. The forecasting results show that using historical features will result in superior performances. However, the differences in performance are not as pronounced as one would originally expect, making this approach a viable alternative to forecasting household demand when historical consumption time series are not available.

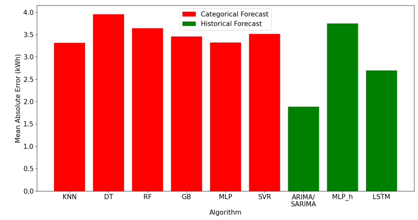
The main limitations of this research stem from data constraints, as obtaining reliable data from publicly available datasets proved challenging. Despite utilizing the REFIT dataset, which offers extensive time-series data and household details, its limited scope, covering only twenty households in the UK, poses potential bias in outcomes. Future research could benefit from accessing more extensive and diverse datasets.



(a)



(b)



(c)

Fig. 7: Mean Absolute Error in the two scenarios: (a) 1-Day average, (b) 7-day average, (c) 30-day average.

Future work should also study feature relevance for predicting household demand since not all features will exhibit the same weight. For instance, appliances with higher consumption will likely have a more pronounced effect in predicting household consumption. Another future research direction would be to explore ensemble-based methods such as stacked generalization to enhance the forecasting models' predictive accuracy further [24]. Furthermore, by leveraging ensembles, it should also be possible to define prediction intervals to quantify the uncertainty associated with the model's forecasting results.

Acknowledgement

This research was funded by FCT project 10.54499/2022.15771.MIT through national funds. The authors further acknowledge FCT for projects 10.54499/LA/P/0083/2020; 10.54499/UIDP/50009/2020 & 10.54499/UIDB/50009/2020, and grants 2020.08075.BD (JG) & CEECIND/01179/2017 (LP).

References

[1] P. H. Bertoldi, *B. An econometric model to access residential electricity savings in the EU. In Proceedings of the 7th International Conference EEDAL'2013 Energy Efficiency in Domestic Appliances and Lighting*, 2013.

[2] J. E. Carlisle, S. L. Kane, D. Solan, and J. C. Joe, "Support for solar energy: Examining sense of place and utility-scale development in california," *Energy Research & Social Science*, vol. 3, pp. 124–130, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S221462961400084X

[3] H. Guda and U. Aliyu, "Design of a stand-alone photovoltaic system for a residence in bauchi," *International Journal of Engineering and Technology*, vol. 5, no. 1, pp. 34–44, 2015.

[4] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3125–3148, 2019.

[5] "How many smart meters are installed in the united states, and who has them?" https://www.eia.gov/tools/faqs/faq.php?id=108&t=3, accessed: 2022-11-29.

[6] B. Völker, A. Reinhardt, A. Faustine, and L. Pereira, "Watt's up at home? smart meter data analytics from a consumer-centric perspective," *Energies*, vol. 14, no. 3, 2021. [Online]. Available: https://www.mdpi.com/1996-1073/14/3/719

[7] Z. Abreu and L. Pereira, "Privacy protection in smart meters using homomorphic encryption: An overview," *WIREs Data Mining and Knowledge Discovery*, vol. 12, no. 4, p. e1469, 2022.

[8] C. S. Bojer, "Understanding machine learning-based forecasting methods: A decomposition framework and research opportunities," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1555–1561, Oct. 2022.

[9] O. Guerra Santin, "Behavioural patterns and user profiles related to energy consumption for heating," *Energy and Buildings*, vol. 43, no. 10, pp. 2662–2672, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778811002738

[10] K. Amasyali and N. M. El-Gohary, "A review of data-driven building energy consumption prediction studies," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1192–1205, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364032117306093

[11] W. van Raaij and T. M. Verhallen, "Patterns of residential energy behavior," *Journal of Economic Psychology*, vol. 4, no. 1, pp. 85–106, 1983. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0167487083900478

[12] J. Paauw, B. Roossien, M. Aries, and O. G. Santin, *Energy Pattern Generator Understanding the effect of user behaviour on energy systems*. Petten: ECN, 2009.

[13] E. de Groot, M. Spiekman, and I. Opstelten, "Dutch research into user behaviour in relation to energy use of residences," *Natural gas*, p. 2, 2006.

[14] W. Poortinga, L. Steg, C. Vlek, and G. Wiersma, "Household preferences for energy-saving measures: A conjoint analysis," *Journal of Economic Psychology*, vol. 24, no. 1, pp. 49–64, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016748700200154X

[15] S. Leth-Petersen and M. Togeby, "Demand for space heating in apartment blocks: measuring effects of policy measures aiming at reducing energy consumption," *Energy Economics*, vol. 23, no. 4, pp. 387–403, 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140988300000785

[16] M. Lenzen, M. Wier, C. Cohen, H. Hayami, S. Pachauri, and R. Schaeffer, "A comparative multivariate analysis of household energy requirements in australia, brazil, denmark, india and japan," *Energy*, vol. 31, no. 2, pp. 181–207, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544205000113

[17] J. Góis, L. Pereira, and N. Nunes, "A Data-Centric Analysis of the Impact of Non-Electric Data on the Performance of Load Disaggregation Algorithms," *Sensors*, vol. 22, no. 18, p. 6914, Jan. 2022.

[18] D. Murray, L. Stankovic, and V. Stankovic, "An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study," vol. 4, p. 160122.

[19] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization." *Journal of machine learning research*, vol. 13, no. 2, 2012.

[20] T.-T. Wong and P.-Y. Yeh, "Reliable accuracy estimates from k-fold cross validation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1586–1594, 2020.

[21] "Cross-validation: K fold vs monte carlo," https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b, accessed: 2023-07-06.

[22] L. J. Tashman, "Out-of-sample tests of forecasting accuracy: an analysis and review," *International journal of forecasting*, vol. 16, no. 4, pp. 437–450, 2000.

[23] S. Rosset, C. Perlich, and B. Zadrozny, "Ranking-based evaluation of regression models," in *Fifth IEEE International Conference on Data Mining (ICDM'05)*, 2005, pp. 8 pp.–.

[24] M. P. Sesmero, A. I. Ledezma, and A. Sanchis, "Generating ensembles of heterogeneous classifiers using Stacked Generalization," *WIREs Data Mining and Knowledge Discovery*, vol. 5, no. 1, pp. 21–34, 2015.