Dataset Justification & Literature Review

Energy Consumption Forecasting: Multi-Model Time Series Analysis

Date: October 7, 2025
Author: Lahiru Manulanka Munasinghe
Project: Machine Learning Time Series for Energy Consumption Forecasting: Multi-Model Time Series Analysis

1) Dataset Description and Justification

We use the UCI Individual Household Electric Power Consumption (IHEPC) dataset that Hebrail and Bérard (2012) put together. The data shows minute-level electricity use from one household in Sceaux, France (a suburb of Paris) from December 16, 2006, at 17:24:00 to November 26, 2010, at 21:02:00. This is about 1,441 days. The dataset has 2,075,259 rows and 8 columns at its native resolution. There is a small amount of missing data in the original file (about 1.2518% overall), which is normal for high-frequency telemetry and can be handled with normal preprocessing. We group the series into daily granularity for modelling to cut down on noise, make sure they match day-ahead operational decisions, and keep training efficient while keeping important temporal patterns.

The main variables show both the total load and its parts: Global_active_power (our target, in kW), Global_reactive_power (kW), Voltage (V), Global_intensity (A), and three sub-metering channels (Wh) for different groups of appliances. We get 1,442 days by 7 numeric columns after daily aggregation. Then, feature engineering adds 46 columns, and we keep a curated set of 44 model features. The exploratory analysis shows that there is a strong trend (decomposition trend strength ≈ 0.6441) and a weekly seasonal signal (seasonal strength ≈ 0.1202). This matches the cycles of human activity. The daily target's stationarity tests are good: The ADF test shows that the data is stationary ($p \approx 0.0046$), and the KPSS test does not reject stationarity ($p \approx 0.10$). We still add differencing features to protect against regime changes and to help models that work better with stationary inputs. Lastly, we divide the data into 1,292 training days, 60 validation days, and 60 test days in chronological order to help with strong model selection and fair evaluation.

This dataset works well for what we need. It provides high-resolution, real-world consumption data with multi-year coverage and extensive covariates, facilitating both univariate and multivariate approaches. Since it is a public benchmark, the results can be repeated and compared to previous work. Most importantly, daily aggregation makes the data fit with day-ahead planning issues in homes. energy management, such as demand response, tariff optimization, and storage scheduling.

2) Forecasting Target and Horizon

We want to predict the daily Global_active_power (kW), which is the daily average of the minute-level measurements. The main time frame for making decisions is one day ahead. This choice is in line with what is done in the industry and helps with planning and utility operations at the household level. For example, decisions about buying energy, responding to dynamic pricing, and scheduling appliances are usually made a day in advance. Day-ahead is a good balance between being able to act on forecasts and being able to trust them. It keeps regular calendar effects while avoiding the error accumulation that can happen over longer time periods.
We also look at secondary, longer windows of 7 and 30 days in addition to the main horizon. These windows help us understand multi-step behaviour and plan for the medium term. These longer time frames are usually more difficult and have more uncertainty, but they give us useful information about how strong and variable a model is over time. Our setup and notebooks are set up to work with both the operational day-ahead use case and these extra analyses.

3) Literature Survey (Public Dataset and Related Work)

The IHEPC dataset is extensively utilised in residential load forecasting research, and the broader problem domain possesses a well-established body of literature. Here, we summarise a few studies that together motivate our choices for features, models, and evaluation design.
Hebrail and Bérard (2012) presented the dataset through the UCI Machine Learning Repository, detailing the data collection methodology, essential variables, and quality considerations. Their contribution made IHEPC a standard for the community and brought attention to common problems like variability, seasonal effects, and non-trivial missingness, which we deal with directly in our pipeline.

Kim and Cho (2019) investigated hybrid deep learning architectures for residential energy forecasting, integrating convolutional layers with LSTMs to identify both localised patterns and extended temporal dependencies. Their CNN-LSTM outperformed ARIMA, SVR, and plain LSTM baselines, with a MAPE of about 4.76% for hourly predictions. This shows how useful non-linear modelling and sub-metering features can be.

Torres et al. (2021) examined deep learning methodologies for time series forecasting, encompassing RNN/LSTM/GRU families, CNN-based techniques, and attention/Transformer architectures. The survey makes it clear that LSTM variants work well for short- to medium-term horizons, while attention mechanisms and Transformers can be helpful for longer sequences. It also emphasises the ongoing significance of meticulous feature engineering and the advantages of ensembling for enhanced robustness.

Kong et al. (2019) concentrated on short-term residential load forecasting utilising LSTM models enhanced by attention mechanisms and external covariates, including weather and calendar variables. Their findings demonstrate that these supplementary signals significantly enhance accuracy, exhibiting a robust R^2 at short-term horizons and a tolerable decline at 24-hour forecasts. This work strongly encourages us to add calendar features and hints that adding weather to future versions could be beneficial.

Lim et al. (2021) presented the Temporal Fusion Transformer (TFT), a model engineered for interpretable multi-horizon forecasting utilising variable selection networks and attention mechanisms. TFT showed big improvements over baseline LSTMs on different datasets and gave quantile forecasts to help with estimating uncertainty. The architecture's ability to handle mixed covariates and be understood fits with our goals of both accuracy and explainability.
Bouktif et al. (2018) examined LSTMs enhanced through genetic algorithms for hyperparameters and feature selection, revealing significant reductions in RMSE and model complexity. Their findings advocate for systematic exploration of model designs and underscore the significance of feature parsimony, which is reflected in our curated feature list and our focus on reproducible configuration.

These studies prompt our multi-model strategy (statistical baselines, tree ensembles, and deep learning), our feature collection (lags, rolling statistics, cyclical encodings, and calendar indicators), and our assessment framework (consistent splits and a cohesive metric suite). Not all previous studies utilised IHEPC specifically; however, they examined analogous residential load series with comparable resolutions and covariates, rendering their conclusions directly applicable to our context.

4) Positioning and Contribution of This Project

Our contribution is a reproducible, configuration-driven forecasting workflow that goes from raw minute-level data to predictions for the next day through clear preprocessing, feature engineering, and systematic model comparison. The pipeline cleans up and fills in missing values, combines them into daily resolutions, marks outliers, and builds a rich but controlled feature set that includes temporal, cyclical, lagged, rolling, exponentially weighted, and differenced signals. We compare classical baselines with modern machine learning and deep learning methods using a consistent chronological split. We choose the best model based on test RMSE and keep all trained artefacts with metadata for traceability. The default day-ahead horizon meets immediate operational needs, while the optional 7- and 30-day horizons help with planning and stress-testing. Notebooks record all the steps, and markdown files summarise them for easy review and conversion to .docx format. This makes sure that the work can be done again and passed on.References

1. Hebrail, G., & Bérard, A. (2012). Individual household electric power consumption. UCI Machine Learning Repository. DOI: 10.24432/C58K54
2. Kim, T. Y., & Cho, S. B. (2019). Predicting residential energy consumption using CNN-LSTM neural networks. Energy, 182, 72–81.
3. Torres, J. F., Hadjout, D., Sebaa, A., Martínez-Álvarez, F., & Troncoso, A. (2021). Deep learning for time series forecasting: A survey. Big Data, 9(1), 3–21.
4. Kong, W., Dong, Z. Y., Jia, Y., Hill, D. J., Xu, Y., & Zhang, Y. (2019). Short-term residential load forecasting based on LSTM recurrent neural network. IEEE Transactions on Smart Grid, 10(1), 841–851.
5. Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748–1764.
6. Bouktif, S., Fiaz, A., Ouni, A., & Serhani, M. A. (2018). Optimal deep learning LSTM model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. Energies, 11(7), 1636.