Leveraging Anomaly Detection and AutoML for Modelling Residential Measurement Power Traces

Grigore Stamatescu
Automation and Industrial Informatics
University Politehnica of Bucharest
Bucharest, Romania
grigore.stamatescu@upb.ro

Radu Plamanescu
Faculty of Electrical Engineering
University Politehnica of Bucharest
Bucharest, Romania
radu.plamanescu@upb.ro

Mihaela Albu
Faculty of Electrical Engineering
University Politehnica of Bucharest
Bucharest, Romania
mihaela.albu@upb.ro

2023 IEEE 13th International Workshop on Applied Measurements for Power Systems (AMPS) | 979-8-3503-1240-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/AMPS59207.2023.10297201

***Abstract*—Emerging data analytics approaches applied in power systems enable the extraction of relevant information from large scale measurement time series. These can be used to improve system observability and operation. Several open datasets can serve as benchmark that allow comparative results between pre-processing, modelling and evaluation strategies. We present a data analytics application on high reporting rate power measurements from single and multiple residential units. Outlier detection to label anomalies in the time series can be combined with automated machine learning libraries for efficient forecasting and dimensionality reduction. We apply this method for establishing an empirical relation between the occurence rate of outliers in the measurement time series and the forecasting performance. Results provide a discussion on the best, domain-adaptive, parametrisation and modelling options which are highly suited for power system measurements using open source software libraries.**

***Index Terms*—data analytics, anomaly detection, automl, smart meter, power measurements**

I. Introduction

Widespread adoption of smart meters and advanced metering infrastructures (AMI) enable fine grained collection of electrical energy measurement data with high temporal and spatial resolution. Information contained in such datasets can be efficiently processed and extracted for improved control of power systems through state-of-the-art algorithms, models and tools [1]. Online analysis of streaming data can contribute to early warning of power quality conditions and transient behaviours as well as potentially unstable conditions that have a cascading effect on the larger grid environment. The challenge at high reporting rates lays in intelligent selection and dimensionality reduction of the raw data that can selectively include only the relevant features in the analysis.

As electrical measurements from power systems are usually presented as uni- or multi-variate time series, appropriate methods can be used to pre-process and model such datasets. The addition of domain knowledge from power system specialists in the pre-processing and feature engineering stage contributes to a more efficient approach by restricting the experimentation to approaches that are relevant also from an engineering point of view, in relation to the observed physical system. These are directly applicable consumers or the local microgrid in our case.

Various methods for modelling time-series range from basic statistic and econometric models, to conventional machine learning models and end-to-end deep learning pipelines [2] that automate the feature extraction and modelling. Current software frameworks can help automate the testing and experimentation of such models and guide the developer through various parametrisation requirements in order to select the best model that fits a particular physical system and its context defined by the data. Outlier (Anomaly) detection is usually implemented in a preliminary phase in order to focus the processing on the particular features that characterise transient or unforeseen behaviours thereby implicitly achieving dimensionality reduction of the input dataset.

By combining anomaly detection with automated machine learning, in the context of power systems measurements, we aim to improve the data analytics approach for such data and contribute to the real-time operation and control of local microgrids. The presented use case is focused on single and multiple residential units (student dormitories) by using a previously deployed open metering infrastructure that enables data collection and aggregation.

Main contribution are thus summarised:

- An approach to combine outlier detection with automated forecasting frameworks that investigates the prediction performance in relation to the number of outliers in a measurement time series;
- An application of the proposed methodology and associated results discussion, considering various parametrisation options.

The rest of the paper is structured as follows. Section II introduces the context of our work through related approaches that focus on data analytics for applied measurements in power systems. Section III discusses the main methods that were implemented to achieve the improved auto-ml forecasting and dimensionality reduction by outlier detection. The datasets used and their collection is also presented here. Section IV lists in depth the achieved results and provides insight into the software and implementation and parametrisation of the work. Section V concludes the paper with outlook on potential practical value and extension of the approach.

This work was supported by a grant of the Romanian Ministry of Education and Research, UEFISCDI no. PN-III-P4-ID-PCE-2020-2876 ”Advanced Measurement Framework for Emerging Electric Power Systems” (EMERGE).

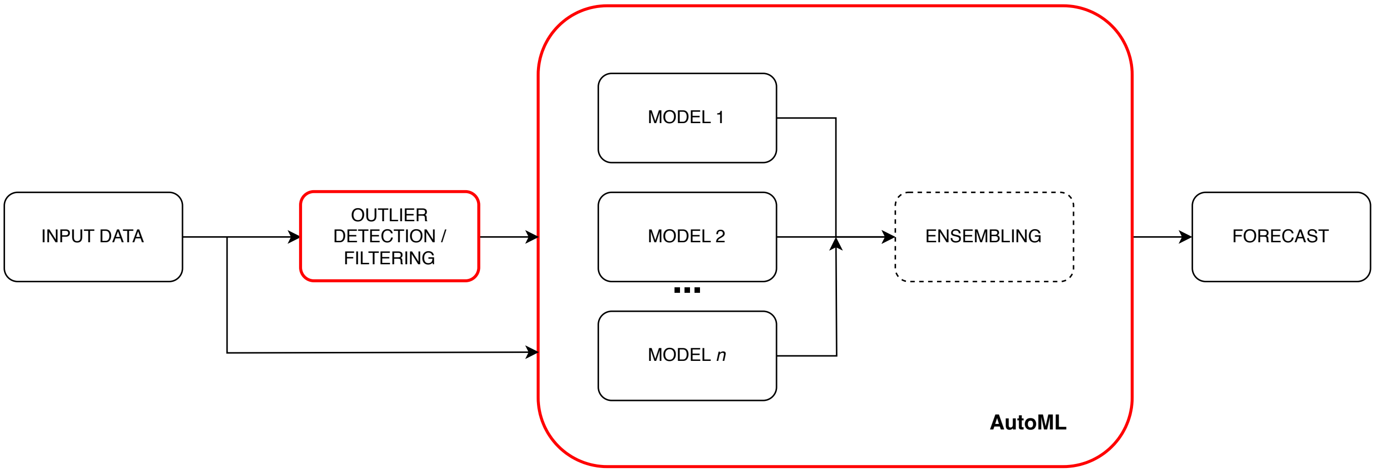

Fig. 1. AutoML pipeline for power measurement processing using outlier detection

II. Related Work

An increasing number of interdisciplinary articles on smart meter data analytics have become available at the intersection of the power systems, measurements and computer science domains. The focus on both the infrastructure and systems to collect, store and analyse such datasets [3] and on the methods and algorithms that are used to efficiently extract relevant information and patterns from the collected data [4]. Embedded and real-time energy analytics [5] can be deployed directly on the smart meter in order to use the local available and increasing computing and communication resources and online decision output at lower sampling periods.

A generic definition of anomalies in power system measurement traces can be seen as ”patterns that deviate from a well defined notion of normal behaviour”. In order to partially mitigate the challenge of accurately identifying and labelling a large number of anomalies in energy and power measurements time series, the authors of [6] propose a method of modelling and generating synthetic anomalies to improve machine learning model training. Such approach can be validated through the statistical properties of the anomalies being considered or through the help of domain experts. Bootstrapping or semi-supervised methods are also available in which a small number of labelled anomalies are available and their characteristics are used to extrapolate and identify a larger number of occurences. A high performance method for time series data analytics (TSAD) based on the Matrix Profile data mining algorithm is presented by [7]. The application concerns online streaming data for which real time results are important through Discord Aware Matrix Profile (DAMP) while the authors argue an analytics throughput of over 3000000Hz on standard hardware.

Automated machine learning (AutoML) pipelines for fast and practical evaluation of classification and prediction performance is being increasingly used in diverse engineering areas. Given a well formulated problem and an initial set of models and hyperparameters, the approach can guide the user to a suitable working solution. A reference is provided in [8] where building energy forecasting is carried out with good quantitative performance. The system is complemented by explainable artificial intelligence (xAI) features that can improve the understanding of the outputs by the end-user, identifying the most important features in the determination of the final result. Such methods can integrare both baseline and state-of-the-art algorithms and combine their outputs, e.g. through ensemble voting schemes, in order to provide robust forecasting outputs over a wider range of input variability.

Previous works such as [9], in which discord labelling through the matrix profile technique is applied, investigated the stability of the detected anomalies under various noise assumptions. In [10] the impact of the reporting rate of the input energy measurement time series on the prediction performance of various deep learning models has been investigated. In [11] multiscale data analytics has been carried out in order to evaluate the performance of various methods at representative and domain specific time scales.

III. Methods and Datasets

A diagram of the proposed system is shown in Figure 1. The main four stages in the data processing pipeline are described:

1) **Input Data**: Power measurement values are read in offline mode from text files containing the readings and associated timestamps or directly, in online mode, by querying the metering infrastructure or through database Application Programming Interfaces (API); the values are stored in structured format e.g. dataframe format, in the development environment for the next stage;
2) **Outlier Detection / Filtering**: The anomaly detection and labelling routing is run on the structured and cleaned data; The datapoints that are identified as outliers are properly marked;
3) **AutoML**: The automated machine learning procedure involves training a number n of distinct parametrised forecasting models on the datasets that include and exclude the outliers from the previous steps; As an optional sub-step, the predictions of the different models

can be combined by computing a weighted average of the model predictions, using the inverse of the Mean Squared Error (MSE) metric as weight; This step allows the quantification of the outlier removal on the forecast accuracy;

4) **Forecast**: The final forecast result is presented to the end-user or provided to the decision and control system for further processing.

A. Theoretical Background

The first step of our application uses the Hampel filter method as proposed by [12]. The method uses a sliding window applied to the measurement time series in which the individual values are compared to the statistical distribution of their neighbors in order to flag and replace the considered outliers in the original time series.

The standard deviation of a data series is computed as $\hat{\sigma} = \sqrt{1/n \sum (x_i - \bar{x})^2}$, with n the number of observations in the window and $\bar{x}$ the window average. The Median Absolute Deviation (MAD) indicator represents the median of the absolute deviations from the median:

$$MAD = median(|X_i - \tilde{X}|) \tag{1}$$

where the median $\tilde{X} = median(X)$. It quantifies the variability of a univariate sample of data and it's considered statistically robust. The MAD is linked to the standard deviation by the formula:

$$\hat{\sigma} = k \cdot MAD \tag{2}$$

where k is a constant scale factor with $k = 1/(\Phi^{-1}(3/4)) \approx 1.4826$ for a normal distribution. The quantile function Φ^{-1} represents the inverse of the cumulative distribution function.

The main input parameters for the Hampel filter are the number of standard deviations and the window size, which can be used to tune the algorithm towards a more restrictive or more lenient detection of outliers. Filtering assumes that once a value does not pass the test and can be considered as an outlier, its value is replaced by the median of its neighbors.

Auomated machine learning increases the efficiency of testing various configurations of machine learning pipelines for engineering applications and can be used to improve and aggregate prediction performance. The approach combines a family of machine learning models and their parameters with a higher level Bayesian optimisation layer which further optimises the hyper-parameters concerned with model selection and evaluation. *auto-sklearn* [13] is a popular example of such tool that employs meta-learning to optimise performance on large datasets using a bandit strategy for budget allocation. The predictions of the generated machine learning pipelines can be combined through robust ensambling.

For time series data, the *auto-ts*[1] package is a suitable alternative to automate time series modelling and forecasting. Multiple pre-processing options, models, ensembling and evaluation metrics are supported in order to adapt the forecasting strategy and improve it through domain expertise (process knowledge). Several other packages have become available that provide a similar approach which includes data pre-processing, feature engineering, hyperparameter optimization, forecasting method selection and forecast ensembling [14].

B. Data

Two residential active power measurement datasets are used to illustrate our approach: an individual housing unit (apartment) from Bucharest, Romania, and a multiple housing unit building (student dormitories) from the campus of the University Politehnica of Bucharest, Romania.

The first dataset is available on IEEE Dataport [15] and includes 1s reporting rate active power measurements collected over a period of several months. The main features are determined by the household appliance consumption patterns, with a few always-on appliances (refrigerator, wifi router) and others which can be used periodically or only in a seldom manner. The data is collect using a dedicated smart meter extension module that interfaces through a communication protocol with the smart meter, collects and stores the readings for further processing. A condensed heat map monthly view of the apartment data for the month of September 2020 is presented in Figure 2.

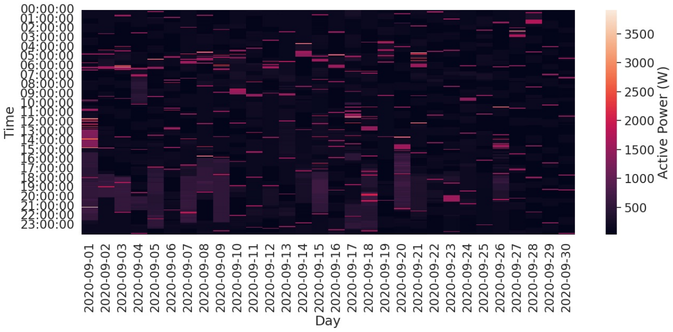

Fig. 2. Monthly active power measurements for a residential unit

Fopr the second dataset we select four days of active power measurement from the dormitories with a 2s reporting rate. Figure 3 illustrates the original measurement time series for these four days. Compared to the apartment, a much higher baseline power consumption is observed given the size and permanent occupancy of the building by several tens of residents. Large seasonal variations are also observed based on the heating/cooling requirements in winter/summer compared to shoulder seasons.

IV. Results

For obtaining the results, implementation has been performed in Python in the Google Colab hosted notebook environment. The data and the Jupyter Notebook code is available on GitHub[2] for replication of the figures and result

[1] https://github.com/winedarksea/AutoTS

[2] https://github.com/grig101/amps23

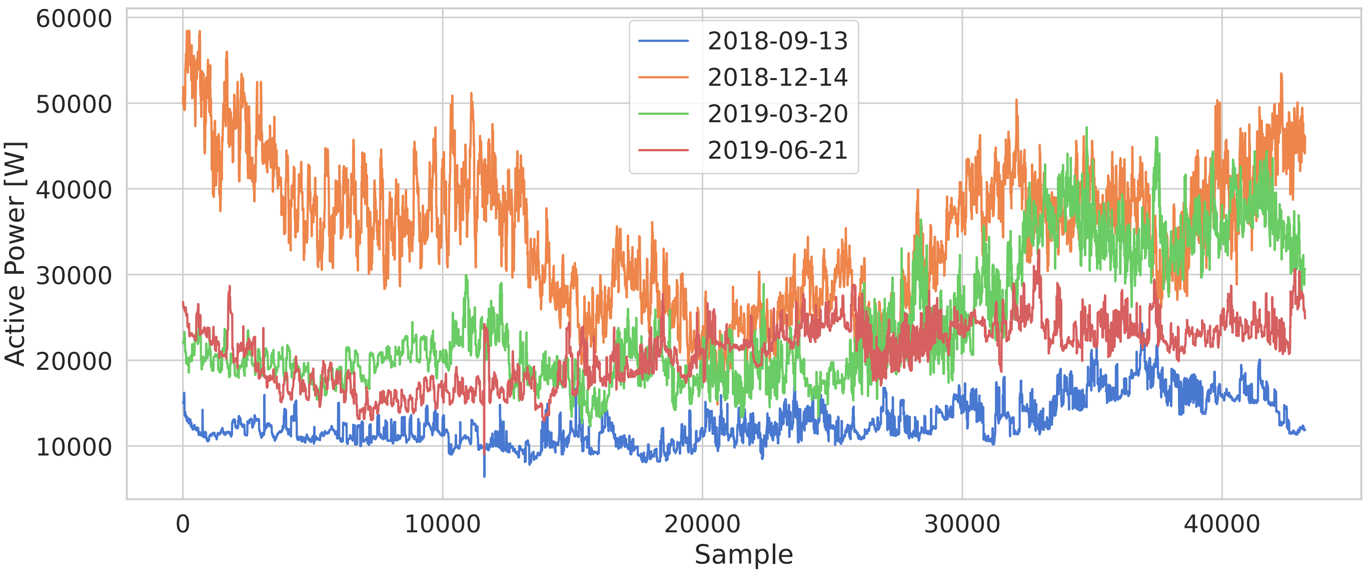

Fig. 3. Daily active power measurements for a multiple residential unit building - student dormitories

tables. Main steps included data import and pre-processing, e.g. timestamp formatting, outlier filtering and automated time series forecasting. An initial example for using Hampel filtering for outlier detection on the dormitory data from September 13th 2018 is introduced in Figure 4. The red points identify the data points in the original timeseries for Figure 3 that have been labeled as outlier using the current configuration of the Hampel filter. The remaining blue line depicts the timeseries with these outliers removed.

The overall outlier rate by using the standard parameters: threshold for number of standard deviations (3) and neighbor window size (15), ranges between 5 and 6 % for the analysed days.

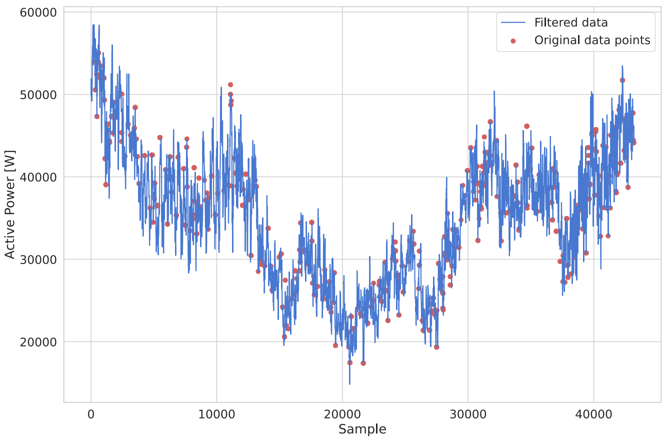

Fig. 4. Outlier detection using Hampel filter

We test the auto time series modelling method on both the original and the Hampel filtered data with the following configuration for the training stage:

```
model = AutoTS(
    forecast_length=3,
    frequency=’infer’,
    model_list=’probabilistic’,
    ensemble=None,
    max_generations=3,
    num_validations=2)
```

where *model_list* denotes the subset of available models in the auto-ts library, with a number of 430 models for the *probabilistic* option. For computational efficiency, we do not use the *ensemble* option while the number of validations is set at 2, for improving model selection with limited penalty on the performance. The sampling rate of the input time series is inferred automatically from the *DateTime* index.

Figure 5 shows the day ahead forecast using the original and filtered data for training at 20s time steps. This allows the qualitative assessment of the prediction performance for the original and filtered - outliers removed, input data, while the quantitative metrics are subsequently introduced in Table I.

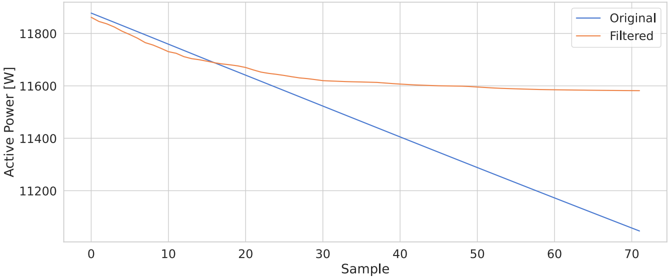

Fig. 5. Day-ahead predictions: original versus filtered data

The metrics used for evaluation during the model selection include the following: Mean Absolute Error (MAE), Symmetric Mean Absolute Percentage loss (sMAPE) and Scaled Pinball Loss (SPL), or Quantile Loss.

MAE only considers the positive variation between the actual and forecasted values is computed as:

$$MAE = \frac{\sum_1^n |y_i - \hat{y}_i|}{n} \tag{3}$$

with y_i the actual value, $\hat{y}_i$ the forecasted value and n the number of samples.

sMAPE is computed as:

$$sMAPE = \frac{1}{n}\sum_1^n \frac{|\hat{y}_i - y_i|}{(|y_i| + |\hat{y}_i|)/2} \cdot 100 \tag{4}$$

and represents a relative performance metric in which the absolute difference between the forecast and the absolute value is divided by half the sum of these values.

SPL function is expressed as:

$$\begin{aligned} SPL(u) = \frac{1}{h}&\frac{\sum_{t=n+1}^{n+h}(y_t - Q_t(u))u1(Q_t(u) \leq y_t)}{1/(n-1)\sum_{t=2}^{n}|y_t - y_{t-1}|} \\ &+\frac{(Q_t(u) - y_t)(1-u)1(Q_t(u) > y_t)}{1/(n-1)\sum_{t=2}^{n}|y_t - y_{t-1}|} \end{aligned} \tag{5}$$

with $Q_t(u)$ the generated forecast for quantile u, h the forecasting horizon and 1 the indicator function.

The values resulting from the AutoML step for the original and filtered data are listed in Table I. The resulting best probabilistic model is the Nonlinear Vector Autoregressive (NVAR) model for this application. The first order NVAR model is expressed as:

$$y_t = f(y_{t-1}, s_t) \tag{6}$$

where $y_t = [y_1(t), ..., y_n(t)]^T$ are the observations and $s_t = [s_1(t), ..., s_n(t)]^T$ are the errors of the process at time t.

TABLE I
AutoML Forecasting Metrics Results

Data	Best Model	MAE [W]	sMAPE [%]	SPL
Original	NVAR	70.2	0.57	0.5
Filtered	NVAR	69.5	0.56	0.485

The reported values correspond to the final, third, validation step which improves the quality of the prediction.

The AutoML procedure covers three categories of models: probabilistic, machine learning and deep learning algorithms. The resulting effect of the filtering yields an improvement in the prediction performance given increased robustness and lower variability of the input data. The characteristics of the determined outliers can be further used to quantify domain-specific variability as additional features. Combining outlier detection with the forecasting step results in improved performance across several domains.

V. Conclusion

We present an approach to combiner outlier detection with automated machine learning pipelines to improve the modelling and forecasting of active power measurements. The presented use case on single and multiple residential units shows good results when comparing the original input and filtered data forecast accuracy. Generalisation of the approach to other datasets will be performed.

References

[1] 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.

[2] E. Eskandarnia, H. Al-Ammal, R. Ksantini, and M. Hammad, "Deep learning techniques for smart meter data analytics: A review," *SN Computer Science*, vol. 3, no. 3, p. 243, 2022.

[3] X. Liu, L. Golab, W. Golab, I. F. Ilyas, and S. Jin, "Smart meter data analytics: Systems, algorithms, and benchmarking," *ACM Trans. Database Syst.*, vol. 42, no. 1, nov 2016. [Online]. Available: https://doi.org/10.1145/3004295

[4] R. Chiosa, M. S. Piscitelli, and A. Capozzoli, "A data analytics-based energy information system (eis) tool to perform meter-level anomaly detection and diagnosis in buildings," *Energies*, vol. 14, no. 1, 2021. [Online]. Available: https://www.mdpi.com/1996-1073/14/1/237

[5] T. Sirojan, S. Lu, B. T. Phung, and E. Ambikairajah, "Embedded edge computing for real-time smart meter data analytics," in *2019 International Conference on Smart Energy Systems and Technologies (SEST)*, 2019, pp. 1–5.

[6] M. Turowski, M. Weber, O. Neumann, B. Heidrich, K. Phipps, H. K. Çakmak, R. Mikut, and V. Hagenmeyer, "Modeling and generating synthetic anomalies for energy and power time series," in *Proceedings of the Thirteenth ACM International Conference on Future Energy Systems*, ser. e-Energy '22. Association for Computing Machinery, 2022, p. 471–484.

[7] Y. Lu, R. Wu, A. Mueen, M. A. Zuluaga, and E. Keogh, "Matrix profile xxiv: Scaling time series anomaly detection to trillions of datapoints and ultra-fast arriving data streams," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1173–1182.

[8] C. Lu, S. Li, S. Reddy Penaka, and T. Olofsson, "Automated machine learning-based framework of heating and cooling load prediction for quick residential building design," *Energy*, vol. 274, p. 127334, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544223007284

[9] G. Stamatescu, R. Plamanescu, I. Ciornei, and M. Albu, "Detection of anomalies in power profiles using data analytics," in *2022 IEEE 12th International Workshop on Applied Measurements for Power Systems (AMPS)*, 2022, pp. 1–6.

[10] G. Stamatescu, I. Ciornei, R. Plamanescu, A.-M. Dumitrescu, and M. Albu, "Reporting interval impact on deep residential energy measurement prediction," in *2021 IEEE 11th International Workshop on Applied Measurements for Power Systems (AMPS)*, 2021, pp. 1–6.

[11] ——, "Multiscale data analytics for residential active power measurements through time series data mining," in *2022 IEEE International Energy Conference (ENERGYCON)*, 2022, pp. 1–6.

[12] T. Brudermueller and M. Kreft, "Smart meter data analytics: Practical use-cases and best practices of machine learning applications for energy data in the residential sector," in *ICLR 2023 Workshop on Tackling Climate Change with Machine Learning*, 2023. [Online]. Available: https://www.climatechange.ai/papers/iclr2023/3

[13] M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, and F. Hutter, "Auto-sklearn 2.0: Hands-free automl via meta-learning," *J. Mach. Learn. Res.*, vol. 23, no. 1, jan 2022.

[14] S. Meisenbacher, M. Turowski, K. Phipps, M. Rätz, D. Müller, V. Hagenmeyer, and R. Mikut, "Review of automated time series forecasting pipelines," *WIREs Data Mining and Knowledge Discovery*, vol. 12, no. 6, p. e1475, 2022.

[15] G. Stamatescu, M. Albu, and M. Sanduleac, "Residential smart meter energy time series: Active power measurements with 1s reporting rate," 2022. [Online]. Available: https://dx.doi.org/10.21227/3yea-xm39