15TH IEA HEAT PUMP CONFERENCE
MAY 26–29, 2026 | HOFBURG
VIENNA 2026
DECARBONISATION THROUGH INNOVATION

Commissioning Oriented Fault Detection in Heat Pump Operation using Machine Learning Algorithms

Parantapa Sawant[a], Nicola von Bülow[a]

[a]University of Applied Sciences Northwestern Switzerland, Hofackerstrasse 30, Basel 4132, Switzerland

Abstract

Heat pumps now play a significant role in Europe's residential heating sector. As deployment continues to expand, there is an increasing need for scalable and commissioning-friendly fault detection (FD) solutions that operate with minimal instrumentation and limited historical data. Existing machine-learning-based FD approaches rely heavily on long process histories or simulation data and require extensive expert configuration, making them unsuitable for early-stage deployment. This paper proposes a lightweight, data-driven FD framework based on short-horizon time-series forecasting using autoregressive–exogenous (ARX) models. The method requires only three consecutive no-fault days for training and incorporates a daily re-initialization loop to mitigate cumulative prediction drift. Deviations between predicted and measured flow-temperature profiles—selected as the primary diagnostic variable due to its universal availability and strong sensitivity to common fault modes—are evaluated using a composite set of complementary residual metrics, including root mean squared error, maximum absolute error, and correlation-based pattern matching.

The framework was tested on seven real-world datasets, with detailed results presented for two representative heat-pump systems: a ground-source heat pump (GSHP) and an air-to-water system. Across both cases, the ARX model achieved strong forecasting accuracy and reliably identified deviations indicative of abnormal behavior, while requiring minimal configuration effort. These results demonstrate the potential of lightweight forecasting-based FD methods as practical commissioning tools and motivate further work toward hybrid approaches that integrate rule-based and data-driven diagnostics for broader real-world deployment.

Keywords: Energy Informatics; Fault-Detection; Time-Series Forecasting

1. Introduction

With annual sales of over 41'000 units and around 23 % of Swiss single-family homes already equipped with a heat pump—despite a recent flattening of growth—the technology is firmly positioned to play a central role in the national heating mix [1]. This growing relevance underscores the need for a scalable and reliable Fault Detection and Diagnosis (FDD) framework to ensure efficient and resilient operation. Many installations lack continuous expert supervision, and undetected faults in early operation can significantly reduce system efficiency and reliability. A commissioning-friendly FDD framework should satisfy the following requirements:

- Minimal additional sensor installation
- Primary reliance on data that is already collected by operators or manufacturers
- Automated evaluation to support human decision-making
- Ability to operate with very limited historical data immediately after commissioning.

Machine Learning (ML) methods for FDD in HP-systems are well documented [2–4]. Most approaches rely on extensive process-history or simulation data describing variables within the heat-pump thermodynamic cycle and employ complex models—such as random-forest regressors and support vector machines—to detect and, in some cases, diagnose faults, with reported accuracies typically ranging between 80 % and 96 %.

However, these methods generally require large amounts of labelled data, often based on high-resolution measurements (e.g. minute-level sampling) collected over periods of several months to a year, which are rarely available during or shortly after commissioning.

To complement the established detailed methods, we propose a novel approach that replaces direct ML classification with a lightweight time-series model enabling early-stage anomaly detection and embedded in a daily re-initialisation loop, thereby reducing cumulative prediction drift. Parameter tuning is fully automated within the framework, substantially decreasing the level of expert intervention typically required in rule-based approaches. Using an *autoregressive–exogenous* (ARX) forecasting model trained on only a few fault-free days, the method achieves a practical balance between scalability and diagnostic robustness immediately after commissioning.

The methodology focuses on the *heat-pump flow temperature*, as this variable is almost universally recorded in heat-pump monitoring systems. Moreover, prior work by Pellella et al. [3], summarising soft and hard faults in residential heat pumps, demonstrates that the flow temperature exhibits strong sensitivity to typical fault types such as fouling, refrigerant overcharge, and leakage, further motivating its use as the primary diagnostic variable.

2. Methodology

The methodology employed in this work adapts a framework previously developed by the authors for domestic solar-thermal systems, where lightweight forecasting-based models were shown to provide reliable diagnostic information even under minimal sensor instrumentation [5].

In the present application, the approach leverages short-horizon, one-day-ahead forecasts of the flow-temperature time series and quantifies deviations between predicted and measured values using a set of complementary error metrics. These prediction errors are aggregated into daily indicators, which are then used to classify each day as either “fault” or “no-fault,” and subsequently compared against rule-based ground-truth labels derived from established operational heuristics. The initial no-fault training periods are selected based on operational plausibility criteria commonly applied during commissioning. These include stable system operation without active error messages, consistent heat demand, absence of known maintenance interventions, and temperature trajectories that follow expected diurnal and load-driven patterns. In this study, the initial no-fault days were identified by an experienced operator using these criteria and served solely as a reference for nominal behaviour.

The overall workflow consists of (i) data preprocessing and selection of no-fault training periods, (ii) construction of ARX features, (iii) model selection and training, (iv) iterative one-day-ahead forecasting using real lagged values, (v) extraction of daily error metrics, and (vi) application of a composite detection rule.

2.1 Selection of Error Metrics

During the development phase, we systematically evaluated a broad range of residual-based indicators commonly used in FDD applications, including mean absolute error, bias computed over sliding windows, residual drift, maximum absolute error, time-over-threshold (ToT), and percentile-based metrics. The evaluation was performed across multiple sensors exhibiting distinct operational characteristics, ranging from stable and moderate variable signals to strongly fault-prone profiles.

The analysis revealed that several metrics were highly correlated and therefore offered limited independent discriminatory power, particularly in cases where the underlying forecasting model exhibited persistent systematic mismatches for specific sensors. ToT and bias-based indicators frequently generated false alarms on sensors with natural mode-switching behavior, while mean absolute error predominantly reflected global amplitude deviations and contributed little to distinguishing fault-specific anomalies.

Through iterative testing we identified three metrics that consistently provided complementary information, were robust across sensors, and aligned well with rule-based ground-truth detection:

- **Spike magnitude (Maximum absolute error, MaxAE)**: Captures short and intense deviations, such as spikes or abrupt control anomalies. This metric is effective for detecting hydraulic disturbances or short-lived control instabilities**.**
- **Global deviation (Root mean squared error, RMSE)**: Represents the average deviation over the entire day and is sensitive to systematic offsets or sustained degradation.

- **Pattern matching (Correlation between predicted and measured daily profiles)**: Identifies pattern mismatches independent of magnitude, especially relevant for detecting malfunctioning plant-states, shifted diurnal profiles, or failures in following ambient-driven dynamics.

2.2 Prediction Model and Feature Construction

To generate the reference predictions required for FD, we tested a set of data-driven forecasting models:

- Classical SARIMAX,
- Linear regression with exogenous inputs,
- Autoregressive models (ARX) with fixed lags,
- Linear models with regularization (Ridge, Lasso),
- Tree-based regressors (Random Forests).

Initial experiments using SARIMAX showed good predictive accuracy, but the computational effort was substantial due to repeated parameter estimation on multi-month high-frequency datasets. For day-by-day operational deployment SARIMAX was judged impractical, particularly when applied to multiple sensors in parallel.

Tree-based models delivered reasonable accuracy but tended to oversmooth short-duration transients, which weakened spike-based detection performance. Moreover, they provided limited interpretability and did not exploit the strong autocorrelation structure typical of hydronic systems.

Based on these comparisons, we selected a regularized linear ARX model, implemented as a Ridge regression with exogenous variables. The chosen model balances:

- low computational effort,
- stable behavior across sensors,
- good prediction quality for both diurnal and transient dynamics, and
- interpretability and reproducibility.

The ARX feature set includes:

- **Lagged sensor values** at the previous time step, one hour, one day, and two days, thereby capturing short-term temporal dependencies as well as daily periodicity.
- **Exogenous variables** representing key boundary conditions, namely *ambient temperature* and the *heat-pump ON/OFF status*.

Model training is performed exclusively on no-fault days to ensure that the forecasting model learns the nominal system behaviour without contamination from faulty operating periods, thereby preventing bias in the residual-based detection metrics.

2.3 One-Day-Ahead Forecasting Procedure

To replicate the expected operational deployment, we implemented a recursive rolling-forward approach. For each forecast day, a recursive one-day-ahead prediction is generated as follows:

- A history window covering the maximum lag is assembled from real measured values immediately prior to the forecast day.
- Predictions for each time-step are computed sequentially, ensuring that predictions do not leak into future days.
- Exogenous inputs for the forecast horizon are taken directly from recorded measurements for testing purposes. Thus, we assume a perfect forecast of the exogenous inputs which should be addressed separately in future studies.
- The resulting predicted time series is aligned with the measured counterpart.

2.4 Daily Metrics and Composite Fault Decision

For each day in the forecast window, the predicted and measured profiles are compared using the metrics defined above. These metrics are then combined through an *OR-rule* as shown in Eq. (1).

$$Fault = (RMSE \geq T_{RMSE}) \ \vee \ (MaxAE \geq T_{MaxAE}) \ \vee \ (corr \leq T_{corr}). \tag{1}$$

This composite rule ensures that the detector reacts either to extreme instantaneous deviations, broad sustained mismatches, or incorrect temporal patterns. The resulting daily flags are finally compared to the flags in the ground truth data to assess detection performance.

2.5 Practical Deployment

The fault-detection algorithm is embedded within a daily re-initialisation loop, as illustrated in **Fehler! Verweisquelle konnte nicht gefunden werden.**. The process begins with the user identifying an initial set of three consecutive no-fault days, which serve as the training dataset for the ARX forecasting model. Using this dataset, the model is fitted and then used to generate a one-day-ahead forecast for the next operational day.

Once the forecast is produced, the measured and predicted values for the target day are compared, and the day is annotated either as a no-fault day or a fault day. This daily annotation is written back into the dataset, thereby enriching the historical record with operator-validated labels. Whenever a day is classified as a no-fault day, it is appended to the training dataset so that the ARX model is continuously updated using the most recent representation of nominal system behaviour. Days classified as fault days are excluded from training but retained in the annotated dataset for later evaluation.

The daily re-initialisation loop is designed to prevent cumulative prediction drift that typically arises when forecasts are chained over multiple days. Instead of propagating model predictions forward in time, the ARX model is re-trained each day using only measured data from the most recent no-fault days. Each daily forecast therefore starts from true observed values, ensuring that model errors do not accumulate across days and that the model remains aligned with evolving system behaviour. This design mirrors practical deployment, where new measurements become available daily and models can be updated incrementally.

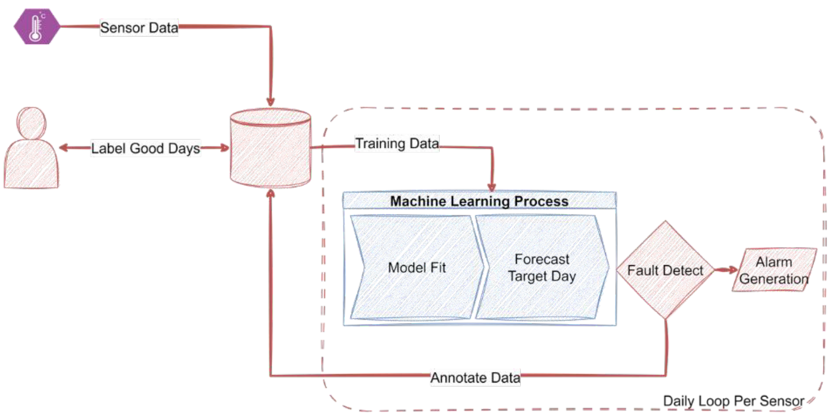


Figure 1. Exemplary framework for implementing the MLA pipeline in fault detection and iterative training-set updates

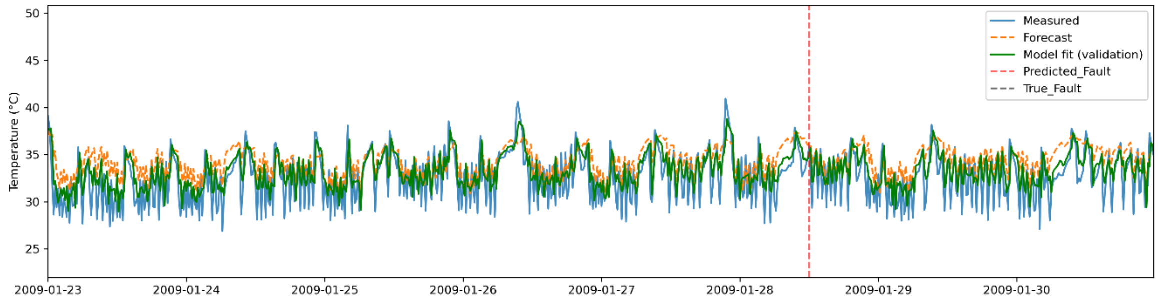
3. Results

The proposed FD framework was evaluated on seven datasets originating from different residential heat-pump installations. Table 1 summarises the forecasting and detection performance across all seven evaluated systems.

Table 1. Summary of results for the seven systems

System	Type	MAE	RMSE	False Positive	False Negative
HP System 1	GSHP Monovalent	1.5	2.0	1	0
HP System 2	Air-Water Parallel	0.35	1.27	6	0
HP System 3	Air-Water Parallel	0.6	1.7	1	0
HP System 4	Air-Water Parallel	0.8	2	5	1
HP System 5	GSHP Parallel	1.17	2.6	6	0
HP System 6	Air-Water Parallel	1	1.8	3	0
HP System 7	Air-Water Parallel	1.3	1.5	5	1

For clarity, we report detailed results for two representative systems: a ground-source heat pump (GSHP) from the *CosyPlace project* (HP System 1) [6] and an air-to-water heat pump (HP System 2) [7]. These systems capture the range of behaviour typically observed in practice, with System 1 integrating a floor-heating distribution system together with domestic-hot-water production and buffer storage, and System 2 including an immersion heater as part of its auxiliary configuration. The results are shown in Figure 2 and Figure 3 below.



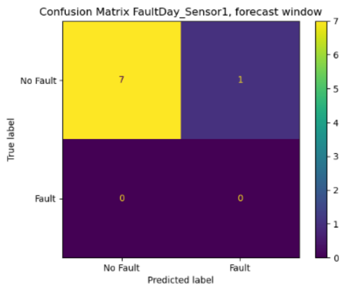
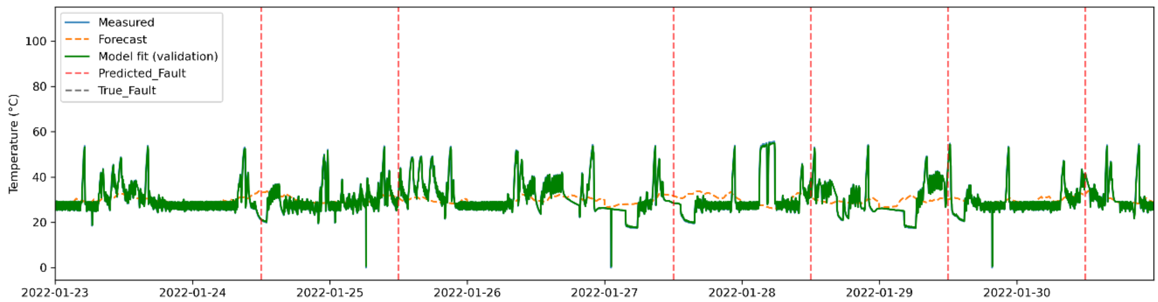


Figure 2. Results for 8-days testing of the ML-Pipeline on winter days in January for HP-System 1

Across both systems, the ARX forecasting model achieved strong predictive accuracy despite being trained on only three no-fault days. For HP System 1, the validation errors were MAE = 1.549 °C and RMSE = 2.039 °C, while HP System 2 yielded even lower errors (MAE = 0.352 °C and RMSE = 1.279 °C). These results confirm that the model approximates the nominal temperature trajectory well and underline the plug-and-play character of the approach across different plant types and operating conditions.



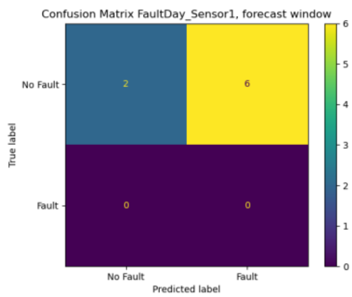


Figure 3. Results for 8-days testing of the ML-Pipeline on winter days in January for HP-System 2

The fault-detection outcomes reflect the contrasting operational characteristics of the two systems. In HP System 1, the forecasts align closely with the measured data, and the algorithm correctly identifies seven no-fault days while generating one false positive. For HP System 2, the more irregular temperature behavior—characterized by rapid transients and mode switching—results in a greater number of threshold exceedances. The algorithm identifies two no-fault days and produces six false positives. Although these elevated false-positive rates point to opportunities for threshold refinement, they remain acceptable in a commissioning context, where conservative detection behavior is preferable and occasional false alarms are not critical.

4. Discussion and Outlook

Overall, the results demonstrate that the proposed method delivers accurate short-horizon forecasts and provides robust anomaly detection with minimal setup effort. The approach remains effective across different heat-pump configurations and operational patterns, which supports its suitability as a commissioning-oriented FD tool. In particular, the plug-and-play nature of the ARX-based pipeline and its reliance on only a few no-fault days make it attractive for practical deployment in settings where historical data and expert configuration effort are limited. The elevated false-positive rate observed for HP System 2 is primarily attributed to its highly transient behaviour and frequent mode switching, which increases sensitivity of residual-based thresholds. Several mitigation strategies are currently under investigation, including adaptive thresholding based on recent residual distributions, persistence criteria requiring repeated exceedances over multiple days, and the integration of additional contextual signals such as control states or auxiliary heater activation. These extensions are expected to significantly reduce false positives while preserving conservative detection behaviour during commissioning.

Looking ahead, several extensions appear promising. First, the simplified forecasting-based approach could be combined with existing rule-based methods or manufacturer-specific diagnostics to form a hybrid FD system that exploits both data-driven residuals and expert knowledge. Such a combination may help reduce false positives and improve interpretability for operators. Second, the framework should be tested on a broader set of real-world datasets covering a wider variety of heat-pump types, control strategies, and fault categories. This would allow a more systematic assessment of its robustness, help refine threshold-selection strategies, and ultimately support its integration into industrial monitoring and building automation platforms.

Acknowledgements

The authors gratefully acknowledge Mr. Andreas Genkinger (INEB–FHNW) for providing the operational data from the CosyPlace Project, which served as the dataset for Heat Pump System 1.

References

[1] Jahres-Statistiken & Marktzahlen –Wärmepumpen Marktentwicklung Schweiz, Fachvereinigung Wärmepumpen Schweiz FWS, Bern, 2025. https://www.fws.ch/statistiken/ (accessed November, 2025).

[2] B. Hofer, F. Wotawa, Detecting Soft Faults in Heat Pumps, in: I. Pill, A. Natan, F. Wotawa (Eds.), 35th Int. Conf. Princ. Diagn. Resilient Syst. DX 2024, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2024: p. 22:1-22:10. https://doi.org/10.4230/OASIcs.DX.2024.22.

[3] F. Pelella, A.F. Passarelli, B. Llopis-Mengual, L. Viscito, E. Navarro-Peris, A.W. Mauro, State-of-the-Art Methodologies for Self-Fault Detection, Diagnosis and Evaluation (FDDE) in Residential Heat Pumps, Energies 18 (2025) 3286. https://doi.org/10.3390/en18133286.

[4] M.S. Mirnaghi, F. Haghighat, Fault detection and diagnosis of large-scale HVAC systems in buildings using data-driven methods: A comprehensive review, Energy Build. 229 (2020) 110492. https://doi.org/10.1016/j.enbuild.2020.110492.

[5] P. Sawant, R. Eismann, Commissioning-Oriented Fault Detection Framework for Building Heating Systems Using SARIMAX Models, in: Proc. Up.-Rhine Artif. Intell. Symp., Proceedings of the Upper-Rhine Artificial Intelligence Symposium, 2024: pp. 59–68. https://doi.org/10.60643/URAI.V2024P59.

[6] A. Genkinger, T. Afjei, R. Dott, A. Witmer, Kuehlung MINERGIE-P Wohngebaeude CosyPlace, 2010. http://hdl.handle.net/11654/12563.

[7] Energy Systems Catapult, EoH Performance DataElectrification of Heat Demonstration Project, 2020-2023Electrification of Heat Demonstration Project: Heat Pump Performance Cleansed Data, 2020-2023, (2024). https://doi.org/10.5255/UKDA-SN-9050-2.