Prototype of a Commissioning-Friendly Fault Detection Tool for Residential Heat Pumps using Machine Learning Algorithms

Parantapa Sawant, Nicola v. Bülow, Institute of Sustainability and Energy in Construction INEB-FHNW

Zusammenfassung

Wärmepumpen sind ein zentrales Element der Wärmewende, doch unerkannte Betriebsfehler können die Systemeffizienz um bis zu 25 % reduzieren. Bestehende Methoden zur Fehlererkennung setzen umfangreiche Betriebsdaten, detaillierte physikalische Modelle oder Expertenwissen voraus — Ressourcen, die unmittelbar nach der Inbetriebnahme typischerweise nicht verfügbar sind. Dieser Beitrag stellt ein schlankes, inbetriebnahmefreundliches Framework zur Fehlererkennung vor, das auf kurzfristiger Zeitreihenprognose mittels eines regularisierten ARX-Modells basiert. Trainiert auf nur drei fehlerfreien Referenztagen, bewertet das Framework tägliche Abweichungen zwischen prognostizierter und gemessener Vorlauftemperatur anhand von drei komplementären Residualmetriken, kombiniert durch eine logische ODER-Regel. Evaluiert auf 12 realen Wärmepumpen-Datensätzen sowie einem synthetischen Datensatz — Luft-Wasser-, Hybrid- und Erdwärmepumpensysteme umfassend — wurden in acht von zwölf Systemen keine Fehler übersehen bei einer mittleren Falschpositivrate von 3,9 Tagen pro System.

Abstract

Heat pumps are central to residential decarbonization, yet undetected operational faults can reduce system efficiency by up to 25%. Existing fault detection methods rely on extensive labeled data, detailed physical models, or specialist configuration — none of which are available immediately after commissioning. This article presents a lightweight, commissioning-friendly fault detection framework based on short-horizon time-series forecasting using a regularized ARX model. Trained on as few as three fault-free reference days, the framework evaluates daily deviations between predicted and measured flow temperature using three complementary residual metrics combined through a logical OR rule. Evaluated on 12 real-world residential heat pump datasets and one synthetic dataset across air-to-water, hybrid, and ground-source configurations, the framework achieved zero missed faults in eight of twelve systems with a mean false positive rate of 3.9 days per system.

Introduction

Heat pumps have emerged as a central technology in modern residential energy systems, where they operate alongside PV, hydronic distribution systems such as underfloor heating, various storages, and an overarching energy management system. In this multi-technology context, the heat pump governs overall system efficiency to a degree that no other single component does, making its reliable and fault-free operation a prerequisite for achieving the energy performance targets that motivate its adoption. Studies have shown that undetected faults can reduce system efficiency by up to 25%, translating directly into elevated energy costs and increased carbon emissions over the operational lifetime of the system [1].

The scale of deployment underscores the urgency of addressing this risk. In Switzerland alone, annual heat pump sales exceed 41'000 units [2], and the technology is firmly positioned as a central pillar of the national decarbonization strategy. As the installed base expands rapidly, so does the population of systems operating without continuous expert supervision — a condition under which performance degradation can go undetected for extended periods, eroding both the efficiency gains and the broader credibility that the technology promises to deliver.

A scalable, reliable, and commissioning-friendly Fault Detection and Diagnosis (FDD) framework is therefore essential. Such a framework must be capable of systematically identifying anomalous behavior early — ideally from the first days of operation — without requiring extensive historical data, specialized sensor installations, or significant expert intervention. This report presents a prototype of such a system, developed and validated on real-world residential heat pump data. Section 1.1 provides the necessary background on FDD concepts and prior work; Section 2 describes the methodology and experimental setup; Section 3 reports the results; and Section 4 discusses the findings and outlines directions for further development.

Background of Fault Detection and Diagnosis in Heat Pump Operation

FDD encompasses two distinct but complementary tasks. Fault detection refers to the recognition that an anomaly is present — that the system is no longer operating within its expected performance envelope. Fault diagnosis extends this by identifying which specific fault has occurred and, where possible, its root cause and in some cases quantify the severity, enabling operators to prioritize and schedule maintenance interventions. The present work focuses primarily on the detection layer, as this represents the most critical and least addressed need at the commissioning stage.

Faults in heat pump systems are broadly classified into *hard faults* and *soft faults*, as illustrated in *Figure 1*. Hard faults — including compressor failure, fan failure, or electronics malfunctions — produce immediate, unambiguous system responses and are generally identified quickly. Soft faults are more consequential in practice: they degrade performance gradually and may persist undetected across entire heating seasons. These are further subdivided into *operational faults* and *installation-related* faults. <u>Installation faults</u>, such as refrigerant overcharge or liquid line restriction, typically manifest during the initial hours of commissioning. <u>Operational faults</u> are more insidious and of greater practical concern: they typically include problems in the refrigerant circuit such condenser and evaporator fouling or refrigerant leakage, but also a broad class of parameterization and configuration errors that are frequently overlooked because they do not trigger error codes and therefore go undetected. Representative examples from residential practice include an incorrectly parameterized heating curve; sensors incorrectly positioned in heating or domestic hot water storage amongst others. The reader is referred to comprehensive reviews [1, 3, 4] of fault taxonomies and their energy impact.

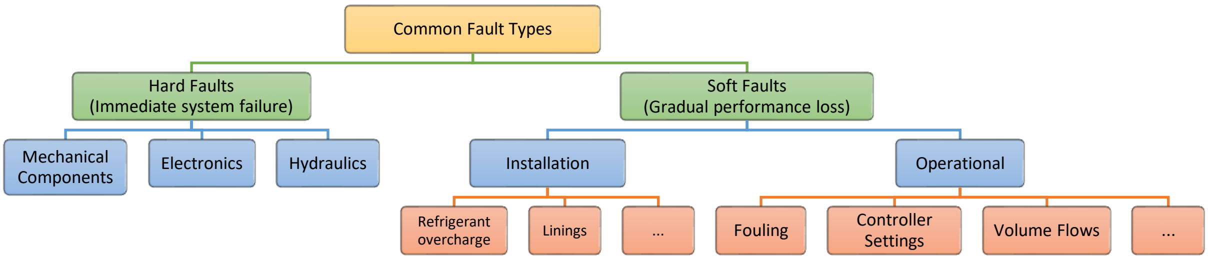


Figure 1 Common fault types in residential heat pump systems (adapted from [1])

FDD methods are generally classified by the type of process knowledge they exploit, and a taxonomy is shown in *Figure 2* below. Model-based approaches construct an explicit system representation using a priori expert knowledge — quantitative approaches through mathematical and simulation-based models, qualitative approaches through causal or heuristic rules. Process-history-based approaches derive diagnostic knowledge entirely from historical operational data without requiring an explicit system model. A practically important subclass is expert systems, which encode diagnostic expertise as rule sets. These are effective for fault diagnosis — particularly for control-related faults — but are inherently incomplete and typically require substantial sensor coverage. Hybrid approaches that combine parity space methods with expert systems exploit their complementary strengths: parity space for detection sensitivity, expert systems for diagnostic specificity. A detailed review of these methodologies and their implementation as products or patents is provided comprehensively by Pelella et al. [1]

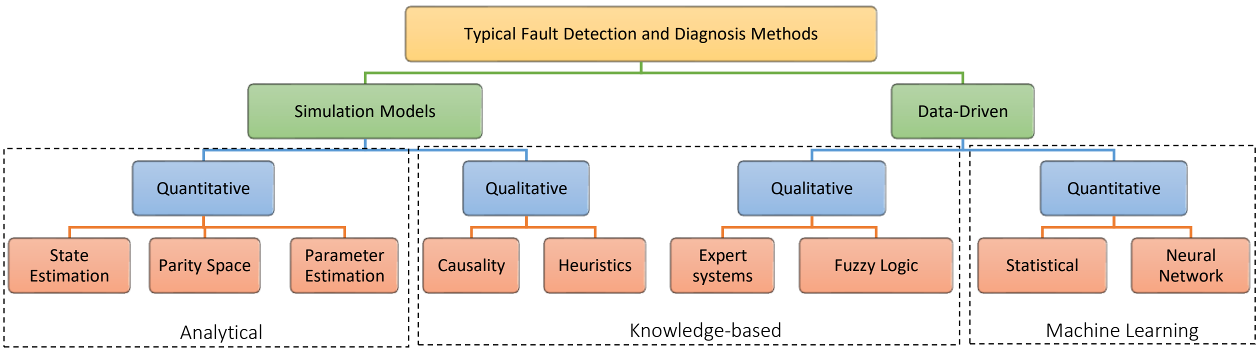


Figure 2 Classification of typical FDD methods in HVAC systems with some examples (adapted from [4])

Within the Swiss research context, following contributions are of relevance:

- At the component diagnostics level, a dissertation project at the ETH developed two complementary fault diagnosis systems for operational monitoring already in the early 2000s. The first employed sequential re-identification of a dynamic grey-box model with fault classification via statistical methods, fuzzy logic, and neural network techniques, designed for minimal additional sensor instrumentation. The second used a steady-state physical model whose parameters carry direct physical meaning, enabling straightforward operator interpretation and requiring minimal training effort [5].
- More recently, the *DIBA-WP project* at HSLU established a systematic basis for digital operational analysis of heat pumps also at a component level, cataloguing 130 fault types of which 27 were selected for detailed investigation. Using simulation-based quantitative evaluation, the most energy- and lifetime-relevant faults were identified — including incorrect heating curve settings, incorrect source pump parameterization in brine/water systems, incorrect switchover temperature to auxiliary electric heating elements, and deteriorating defrost behavior [6].
- At a system diagnosis level, a recent research project at ETH developed and evaluated a set of AI-based algorithms that assess heat pump operation using only 15-minute smart meter electricity readings, without requiring any additional contextual information about the installation. The algorithms enable automated identification of households with heat pumps, differentiation between modulating and non-modulating systems, disaggregation of the heat pump signal from total household consumption, and retrospective estimation of energy savings achievable through operational optimization [7].

Despite this progress, existing FDD methods share several limitations that constrain their applicability to small residential heat pump installations. From a hardware perspective, the most diagnostically informative measurements — such as direct air and water flow rates for identifying heat exchanger fouling, or refrigerant charge sensors for detecting leakage — are typically only available in large-capacity commercial systems where sensor infrastructure is already in place. In residential units, both the physical space and the cost of such instrumentation are prohibitive; in some cases, the cost of a comprehensive sensor suite would approach or exceed the value of the heat pump itself. For residential FDD applications, diagnostic methods must therefore rely exclusively on low-cost, non-intrusive measurements such as temperatures and pressures that are already present in standard installations or can be retrofitted as IoT sensors.

From a data and modelling perspective, machine learning-based FDD approaches have demonstrated strong detection and diagnosis performance, with reported accuracies typically in the range of 80–96% [4, 8, 9]. However, these results are generally achieved under conditions that are incompatible with early-stage deployment: they require large volumes of labelled training data collected at high temporal resolution over periods of several months to a year, along with extensive expert effort to configure and tune models across different installation types. Expert system approaches face a complementary challenge — while they require less data, they demand significant domain knowledge to define rule sets and thresholds, and their scalability across the diverse landscape of residential heat pump configurations is limited. Neither class of method is well-suited to the period immediately following commissioning, when historical data is absent, system behavior is still stabilizing, and the need for diagnostic capability is arguably greatest.

Towards a Commissioning-Friendly Approach

The limitations described above motivate a fundamentally different design philosophy. A commissioning-friendly FDD framework for residential heat pumps should satisfy the following criteria:

- it must operate with very limited historical data immediately after installation;
- it must rely primarily on measurements already collected by operators or manufacturers, with any additional sensor requirements kept to a minimum and limited to low-cost, non-intrusive instrumentation such as temperatures;
- its evaluation must be sufficiently automated to support human decision-making without demanding continuous expert involvement; and
- it must scale readily across different heat pump types and hydraulic configurations.

The core physical intuition behind the proposed approach is that thermal systems exhibit strong day-to-day regularity once they are commissioned and in stable operation. Under normal conditions, the thermal response of a heat pump on any given day closely resembles that of preceding days with comparable boundary conditions — deviations from this pattern are therefore diagnostically informative. Systematic changes in system dynamics do occur, but they follow predictable drivers such as the weekday-weekend cycle or seasonal transitions in outdoor temperature. If nominal operation on a set of verified reference days is known, any significant and persistent departure from that baseline on subsequent days warrants further investigation or classification as a fault.

This observation motivates a lightweight, forecasting-based detection framework. Rather than training a classification model on an extensive labeled fault history, the proposed method establishes a reference model of nominal system dynamics from as few as three consecutive reference days immediately after commissioning. Deviations between predicted and measured output — evaluated using a set of complementary residual metrics — form the basis for daily fault detection decisions. Parameter estimation is fully automated, eliminating the need for expert configuration, and a daily re-initialization loop prevents prediction drift from accumulating over time. The method is deployable from the first days of operation, imposes no additional instrumentation burden, and is designed to support — rather than replace — human judgment in the commissioning process.

Methodology

The fault detection framework presented in this article builds on a forecasting-based diagnostic approach previously developed by the authors for domestic solar-thermal systems, where lightweight time-series models were shown to deliver reliable diagnostic information even under minimal sensor instrumentation [10]. The present work adapts and extends this framework to residential heat pump systems, exploiting the same core principle — that deviations between a short-horizon forecast of nominal system behavior and actual measured values can serve as a practical and scalable fault indicator — while addressing the specific operational characteristics and data availability constraints of heat pump installations.

The overall workflow consists of six steps: data preprocessing and selection of reference training periods; construction of forecasting model features; model selection and training; iterative one-day-ahead forecasting using real lagged values; extraction of daily residual metrics; and application of a composite detection rule. Each of these steps is described in the subsections that follow.

Practical Deployment of Fault Detection Framework

The framework operates as a daily pipeline that alternates between model updating and fault assessment, as illustrated in *Figure 3*. Deployment begins at commissioning, when an operator or technician identifies an initial set of three consecutive days during which the system is known to be operating correctly. These reference days form the initial training dataset for the forecasting model and define the baseline of nominal system dynamics against which all subsequent operation is assessed.

The selection of reference days is based on operational plausibility criteria that are commonly applied during commissioning and do not require specialist measurement analysis. Suitable reference days are characterized by stable operation without active error messages, consistent heat demand, absence of known maintenance interventions, and flow temperature trajectories that follow expected diurnal and load-driven patterns. In the validation study underlying this article, the initial reference days were identified by an experienced operator applying these criteria; in a fully automated deployment, this step could be supported by simple rule-based screening of the available measurement record.

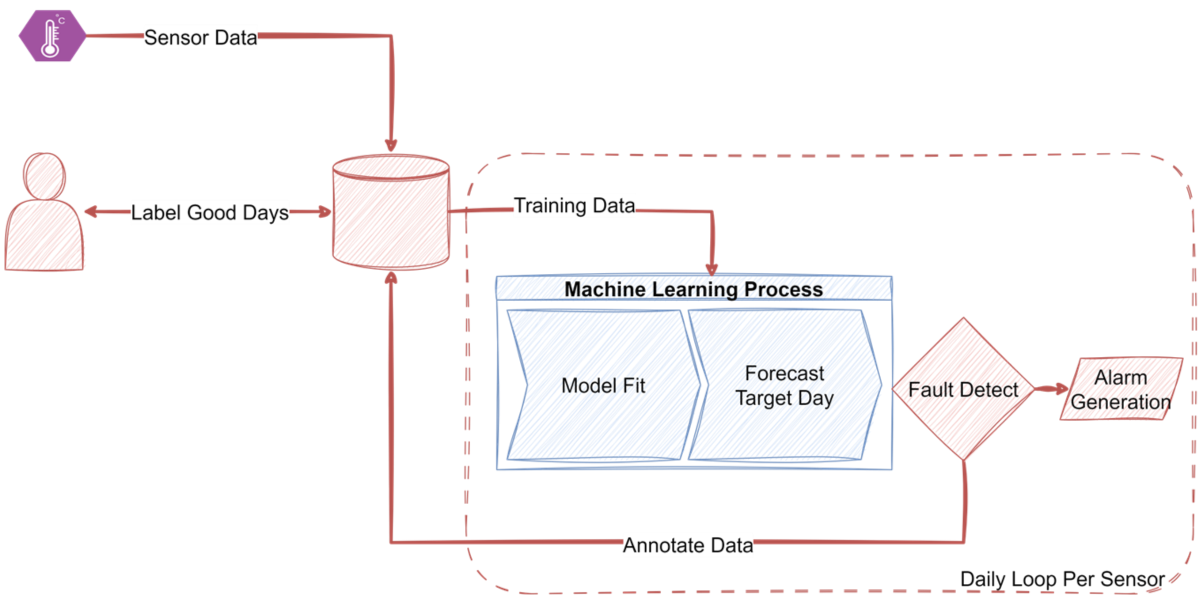


Figure 3 Exemplary framework for implementing the MLA pipeline in fault detection and iterative training-set updates

Once the reference dataset is established, the framework executes the following steps on a daily basis. The forecasting model is fitted to the available reference data and used to generate a one-day-ahead prediction for the target day. The predicted and measured profiles for that day are then compared using the residual metrics described earlier, and the day is classified as either a reference day or a fault day. This classification is written back into the dataset: reference days are appended to the training set, continuously updating the model's representation of nominal dynamics; fault days are excluded from training but retained in the annotated record for later review and potential expert follow-up.

The daily re-initialization loop is a central design feature of the framework. Rather than propagating forecasts forward in a chained sequence — which causes prediction errors to accumulate progressively — the model is re-fitted each day using only measured data from the most recently verified reference days. Each daily forecast therefore starts with true observed values, ensuring that model errors do not compound over time and that the reference model remains aligned with the system's evolving operating conditions. This architecture mirrors practical deployment scenarios in which new measurements become available daily and models can be updated incrementally without manual intervention.

Time-Series Forecasting Model

The primary diagnostic variable used in this framework is the heat pump supply flow temperature. This choice is motivated by two complementary considerations. First, the flow temperature is among the most universally available measurements in residential heat pump monitoring systems, recorded as standard by most manufacturers and operators without requiring additional instrumentation. Second, as documented in the prior literature, the flow temperature exhibits strong sensitivity to the fault modes most encountered in residential installations — including heat exchanger fouling, refrigerant leakage, and hydraulic disturbances — making it a practically informative diagnostic signal across a wide range of fault types [1].

Several candidate forecasting models were evaluated during the development phase, including classical seasonal autoregressive models with exogenous inputs (SARIMAX), standard linear regression with exogenous variables, autoregressive models with exogenous inputs (ARX) with fixed lags, regularized linear models (Ridge, Lasso), and tree-based regressors (Random Forests).

SARIMAX demonstrated competitive predictive accuracy but imposed substantial computational overhead due to repeated parameter estimation on high-frequency multi-month datasets. For day-by-day operational deployment — particularly when applied to multiple sensors in parallel — this computational cost was judged impractical. Tree-based models delivered reasonable accuracy but tended to over-smooth short-duration transients, reducing their sensitivity to spike-based residual metrics. They also offered limited interpretability and did not naturally exploit the strong autocorrelation structure characteristic of hydronic heating systems.

Based on this evaluation, a regularized linear ARX model implemented as a Ridge regression with exogenous inputs was selected. This model achieves a deliberate balance between low computational effort, stable performance across sensors with different operational characteristics, adequate prediction quality for both diurnal and transient dynamics, and sufficient interpretability to support threshold refinement and result communication in practice.

The ARX feature set is constructed to capture both the short-term temporal dynamics and the primary boundary conditions of the system. Lagged flow temperature values at four time offsets — the previous time step, one hour prior, one day prior, and two days prior — are included as autoregressive inputs.

These lags provide the model with information about both immediate thermal inertia and the daily periodicity of heating demand. Two exogenous inputs are incorporated: the ambient outdoor temperature, which is the primary driver of heating load, and the heat pump on/off status, which directly conditions the flow temperature profile. Model training is performed exclusively on verified reference days to ensure that the learned representation of nominal dynamics is not contaminated by fault periods, which would bias the residual-based detection metrics.

Predictions are generated using a recursive rolling-forward procedure that replicates the expected operational deployment. For each forecast day, a history window covering the maximum lag is assembled from real measured values immediately prior to the forecast period. Predictions for each time step are then computed sequentially, using the ARX model with the most recently available lagged values. Exogenous inputs for the forecast horizon are taken from recorded measurements, which in the current implementation assumes perfect knowledge of boundary conditions — an assumption that is conservative in practice. The resulting predicted time series is aligned with the measured counterpart for residual computation.

Residual Metrics and Composite Fault Detection

A single residual metric is rarely sufficient for reliable fault detection across the range of fault modes encountered in residential heat pump operation. Different fault types produce characteristically different signatures: some manifest as brief, intense deviations; others as sustained but moderate offsets; and others as a gradual deterioration in the temporal structure of the flow temperature profile that magnitude-based metrics alone would not capture.

During the development phase, a broad candidate set of residual metrics was systematically evaluated across multiple sensors with distinct operational characteristics, ranging from stable moderate signals to strongly fault-prone profiles. The candidate set included mean absolute error, bias computed over sliding windows, residual drift, maximum absolute error, time-over-threshold (ToT), and percentile-based metrics. Several of these were found to be highly correlated, offering limited independent discriminatory power. ToT and bias-based indicators generated frequent false alarms on sensors exhibiting natural mode-switching behavior, while mean absolute error predominantly reflected global amplitude deviations without distinguishing fault-specific anomalies from normal variability.

Three metrics were retained on the basis of their complementarity, robustness across sensor types, and alignment with fault labels derived from established operational heuristics:

- **Spike magnitude (MaxAE):** The maximum absolute error captures short-lived but intense deviations, such as spikes produced by hydraulic transients or abrupt control anomalies. This metric is sensitive to fault modes that manifest as isolated events rather than sustained drifts.
- **Global deviation (RMSE):** The root mean squared error quantifies the average deviation over the full forecast day and is sensitive to systematic offsets or progressive degradation that persists across many time steps without necessarily producing extreme individual values.
- **Pattern matching (Pearson correlation):** The correlation between the predicted and measured daily profiles assesses the degree of temporal correspondence independent of magnitude. This metric is particularly relevant for detecting faults that alter the diurnal structure

of the flow temperature — such as control failures, mode-switching anomalies, or loss of sensitivity to ambient temperature dynamics — without producing large absolute residuals.

For each forecast day, the three metrics are computed from the predicted and measured profiles and evaluated against empirically determined thresholds. The thresholds T_{RMSE}, T_{MaxAE}, and T_{corr} are derived from the residual distributions observed during the reference period and are set automatically within the framework. These metrics are then combined through an *OR-rule* as shown in Eq. (1).

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

This formulation ensures that the detection rule responds to any of the three characteristic fault signatures independently. A day is therefore classified as a fault day if it exhibits either an extreme instantaneous deviation, a broad sustained mismatch, or an incorrect temporal pattern — without requiring simultaneous exceedance of multiple thresholds. The resulting daily fault flags are compared against ground-truth labels derived from operational heuristics to assess detection performance.

Data and Experimental Setup

The framework was evaluated using data from the Electrification of Heat (EoH) Demonstration Project [11], a large-scale field study conducted in the United Kingdom in which approximately 700 residential heat pump installations were monitored at high temporal resolution. From this dataset, 12 systems were selected to represent a broad range of installation characteristics, including variation in system type (air-to-water, ground-to-water, and hybrid configurations), rated thermal capacity (4–16 kW), and manufacturer. Selection followed a pairwise pairing principle, matching identical or closely related models with markedly different mean flow temperatures and seasonal performance factors, thereby allowing installation-specific influences to be separated from operational anomalies. The selection additionally includes the highest and lowest observed across the full dataset, probing the detection boundaries of the framework across the complete range of observed system behavior.

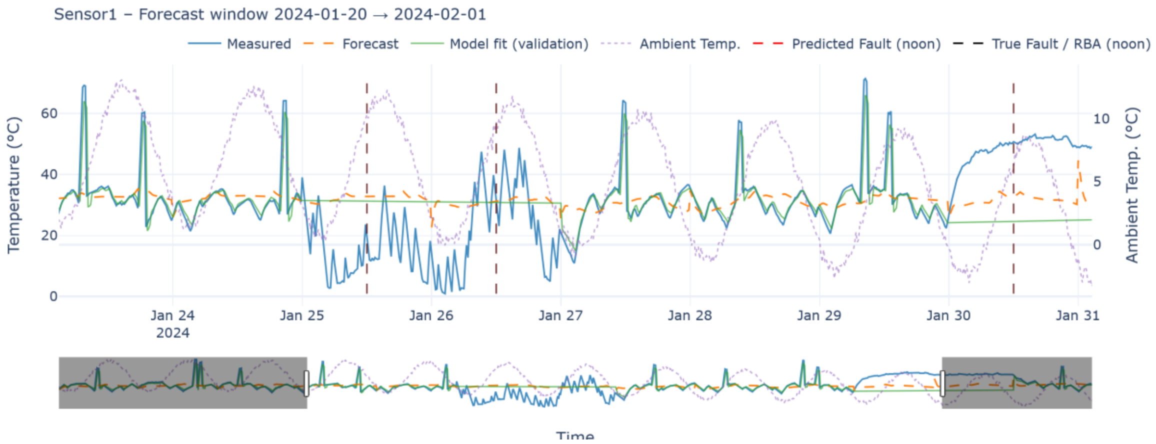
A fundamental limitation of the EoH dataset for FDD purposes is the absence of fault labels. No annotations indicate whether individual days represent normal or anomalous operation, and insufficient information about plant configuration and control parameterization was available to establish ground-truth classifications through expert assessment. A synthetic dataset was therefore generated using the real recordings as a statistical foundation, producing physically plausible flow temperature time series for both normal and faulty operating conditions at 15-minute resolution over 20-day windows. Normal operation was calibrated by linear regression on data from two monitored EoH installations, yielding a heating curve of the form, with compressor cycling modelled as a two-state machine fitted to the observed on/off duration statistics. Two fault types were superimposed on this baseline: a cycling fault, calibrated from a third installation exhibiting behavior consistent with an oversized heat pump operating against insufficient buffer volume and reproducing the characteristic high-frequency flow temperature overshoots above the heating curve setpoint; and a heating curve fault, simulated as a sustained constant flow temperature of approximately 50°C largely decoupled from ambient conditions, consistent with a misconfigured or overridden controller setpoint.

Each synthetic sequence contains a binary ground-truth fault label encoded at noon on each fault day, enabling unambiguous evaluation of detection performance. *Figure 4* illustrates a representative 20-day synthetic sequence showing all three operating regimes — fault-free operation, cycling fault, and heating curve fault — together with zoomed views of each regime and a scatter plot of flow temperature against ambient temperature by operating type, confirming the physical plausibility and mutual distinguishability of the generated signals.

Figure 4 Representative synthetic dataset for a single heat pump instance (DEMO0001) over a 20-day period at 15-minute resolution. Top panel: full time series of flow temperature, ambient temperature, and heating curve set-point, with fault periods shaded. Middle left: zoomed view of fault-free operation on days 4–5, showing realistic compressor cycling. Middle right: zoomed view of the cycling fault on days 11–12, with flow temperature overshooting the heating curve setpoint. Lower left: zoomed view of the heating curve fault on days 16–17, with flow temperature decoupled from ambient conditions. Lower right: flow temperature versus ambient temperature scatter plot by operating regime.

Results

Figure 5 shows the forecast output for the flow temperature (Sensor 1) over the evaluation window from 20 January to 1 February 2024 in the synthetic data set. During the initial reference period, the ARX model fit closely follows the measured flow temperature, capturing both the diurnal cycling pattern and the influence of ambient temperature on the heating curve setpoint. In the forecast window, the model maintains good predictive accuracy during fault-free days, where the forecast and measured profiles remain closely aligned. Two distinct fault periods are visible: a cycling fault characterized by high-frequency, high-amplitude oscillations in the measured flow temperature well below the model forecast, and a heating curve fault in the final days of the window where the measured temperature decouples from the ambient-driven forecast and stabilizes at an elevated level. Both fault periods are correctly identified by the framework, as indicated by the predicted fault markers at noon on the respective days.



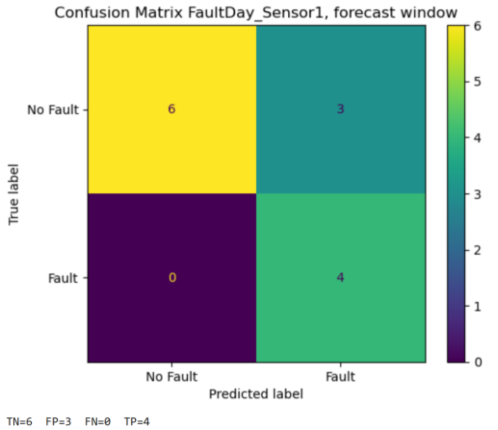


Figure 5 Fault detection results on the synthetic data set

The confusion matrix confirms these observations quantitatively. Across the 13 evaluated days, the framework correctly classified all four fault days (true positive rate = 100%) and correctly identified six of the nine fault-free days (true negative rate = 67%). Three false positives were recorded, meaning that three fault-free days were flagged as anomalous. No fault days were missed (false negatives = 0). In a commissioning context, this outcome is acceptable: the complete absence of missed faults is the primary performance criterion, as undetected faults during early operation carry the greatest risk of sustained efficiency loss. The false positive rate, while non-negligible, reflects the conservative nature of the OR-based composite detection rule and is consistent with the elevated behavioral variability observed in the synthetic dataset during normal operation — particularly the high-amplitude domestic hot water spikes. Strategies to reduce false positives through persistence criteria and adaptive thresholding are discussed in the next.

The proposed fault detection framework was evaluated on 12 synthetic datasets derived from real-world residential heat pump installations, covering air-to-water, air-to-water hybrid, and ground-source configurations with rated capacities ranging from 4 to 16 kW. Figure 6 summarizes forecasting accuracy and detection performance across all systems.

ARX forecasting accuracy was consistently strong across all system types, with MAE values ranging from 0.35 °C to 1.60 °C and RMSE values between 1.27 °C and 2.80 °C, indicating that the model generalizes well without system-specific tuning. Fault detection performance was more variable: false negatives were limited to four systems, each recording a single missed fault day, while false positive counts ranged from 1 to 7 days per system with a mean of 3.9 days. The elevated false positive rates in several systems are primarily attributable to high-amplitude domestic hot water production spikes and mode-switching events during nominally fault-free operation. These results confirm the framework's ability to detect faults reliably across a representative range of residential installations, while identifying threshold refinement and false positive reduction as the primary areas for further development

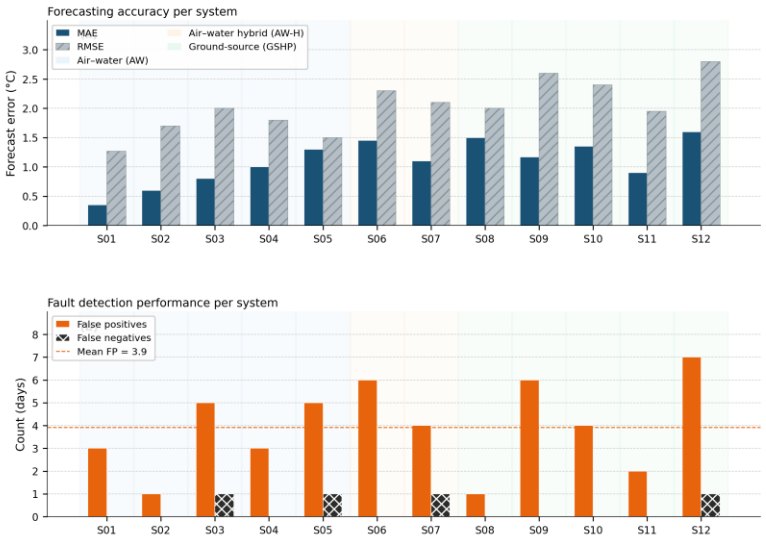


Figure 6 Summary of fault detection performance on 12 different datasets

Discussion and Outlook

The results demonstrate that a lightweight ARX-based forecasting framework can deliver reliable fault detection across a representative range of residential heat pump installations with minimal configuration effort and without requiring labeled fault histories or additional sensor instrumentation. Across all 12 evaluated systems, the framework achieved zero false negatives in eight cases and only a single missed fault day in the remaining four, confirming that the conservative OR-based composite detection rule prioritizes the performance criterion most critical in a commissioning context.

The primary limitation identified is the false positive rate, which averaged 3.9 days per system. This is primarily attributable to high-amplitude domestic hot water production events and frequent mode-switching behavior during nominally fault-free operation, which periodically exceed the residual thresholds established during the reference period. Several mitigation strategies are under investigation: adaptive thresholding based on recent residual distributions, persistence criteria requiring threshold exceedances over multiple consecutive days before a fault is declared, and the incorporation of additional contextual signals such as auxiliary heater activation status. These extensions are expected to reduce false positives substantially while preserving the conservative detection behavior required during early operation.

A further limitation of the current evaluation is the use of synthetic fault data. While the synthetic generation procedure was carefully calibrated to real monitored installations, validation on field datasets with confirmed fault labels remains an essential next step. The assumption of perfect knowledge of exogenous inputs during forecasting — in particular, ambient temperature — also requires relaxation in a production deployment, where forecast boundary conditions would need to be sourced from weather services or estimated from recent measurements.

Looking ahead, several development directions appear promising. Integration with rule-based or manufacturer-specific diagnostics could form a hybrid framework that combines data-driven residual detection with expert knowledge, improving both false positive rejection and fault interpretability for operators. Extension to additional diagnostic variables beyond the flow temperature — such as return temperature differential, compressor on/off statistics, or power consumption — would broaden the detectable fault spectrum. Finally, embedding the framework within a building automation or remote monitoring platform would enable population-level deployment, where aggregated detection outcomes across many installations could support systematic quality assurance of heat pump commissioning at scale.

Bibliography

1. Pelella F, Passarelli AF, Llopis-Mengual B, et al (2025) State-of-the-Art Methodologies for Self-Fault Detection, Diagnosis and Evaluation (FDDE) in Residential Heat Pumps. Energies 18:3286. https://doi.org/10.3390/en18133286

2. (2025) Jahres-Statistiken & Marktzahlen –Wärmepumpen Marktentwicklung Schweiz. Fachvereinigung Wärmepumpen Schweiz FWS, Bern

3. Weigert A (2022) Identification and classification of heat pump problems in the field and their implication for a user-centric problem recognition. Energy Inform 5:70. https://doi.org/10.1186/s42162-022-00250-3

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

5. Zogg D (2002) Fault diagnosis for heat pump systems. IMRT Press, c/o Institut für Mess- und Regeltechnik, ETH Zentrum, Zürich

6. Raphael Agner, Matthias Etterlin, Daniel Troxler, et al (2022) Vermeidung von Effizienzeinbussen von Heizungswärmepumpen durch digitale Betriebsanalyse. SFOE, Horw und Ballwil

7. Brudermueller T, Kreft M, Fleisch E, Staake T (2023) Large-scale monitoring of residential heat pump cycling using smart meter data. Appl Energy 350:121734. https://doi.org/10.1016/j.apenergy.2023.121734

8. Liu Z, Liu Y, Zhang D, et al (2015) Fault diagnosis for a solar assisted heat pump system under incomplete data and expert knowledge. Energy 87:41–48. https://doi.org/10.1016/j.energy.2015.04.090

9. Hofer B, Wotawa F (2024) Detecting Soft Faults in Heat Pumps. In: Pill I, Natan A, Wotawa F (eds) 35th International Conference on Principles of Diagnosis and Resilient Systems (DX 2024). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, p 22:1-22:10

10. Sawant P, Sintzel B, Eismann R, et al Low Cost Monitoring thermischer Solaranlagen mit IoT-Sensor und maschinellem Lernen. https://www.aramis.admin.ch/Texte/?ProjectID=49446. Accessed 11 July 2024

11. Energy Systems Catapult (2024) EoH Performance DataElectrification of Heat Demonstration Project, 2020-2023Electrification of Heat Demonstration Project: Heat Pump Performance Cleansed Data, 2020-2023