GEMMA-FD: Zero-Shot Fault Detection in Heat Pumps Using Multimodal Language Models

Herbert Muehlburger
Institute of Software Engineering and Artificial Intelligence, Graz University of Technology, Austria

Franz Wotawa
Institute of Software Engineering and Artificial Intelligence, Graz University of Technology, Austria

Abstract

Fault detection in heating systems is critical for ensuring energy efficiency and operational reliability. Traditional approaches rely on labeled fault data and expert-defined rules, which are often unavailable or costly to obtain. We introduce GEMMA-FD (*GEMMA for Fault Detection*), a novel zero-shot framework for fault detection in heat pumps that leverages large language models (LLMs) without requiring labeled anomalies or predefined fault signatures. Our method transforms multivariate sensor time series into structured natural language prompts and augments them with visual features, such as line plots of key variables, to facilitate multimodal reasoning. Using GEMMA-3, an open-weight multimodal LLM, we classify heat pump system states as either normal or faulty. Experiments on a real-world heat pump dataset show that GEMMA-FD can identify unseen faults with reasonable precision, although its performance remains lower than a supervised XGBoost baseline trained on the same prompts. Specifically, GEMMA-FD achieves a macro-F1 score of 0.252, compared to 0.69 for XGBoost, underscoring the trade-off between generalization and targeted accuracy. Nevertheless, GEMMA-FD demonstrates the potential of foundation models for interpretable, multilingual fault detection in cyber-physical systems, while highlighting the need for prompt engineering, few-shot augmentation, and multimodal inputs to improve the classification of rare and complex fault types.

2012 ACM Subject Classification Computing methodologies → Machine learning; Computing methodologies → Natural language processing; Software and its engineering → Software testing and debugging; Computing methodologies → Knowledge representation and reasoning; Computing methodologies → Anomaly detection; Computer systems organization → Embedded and cyber-physical systems

Keywords and phrases fault detection, anomaly detection, cyber-physical systems, HVAC, heat pumps, energy systems, large language models, zero-shot learning, open-weight LLMs, interpretable AI, multimodal prompts, smart energy

Digital Object Identifier 10.4230/OASIcs.DX.2025.12

Supplementary Material *Software (Reproducibility package)*: `https://doi.org/10.5281/zenodo.16948128`

Funding This work was partially funded by the FFG project "Scalable Agents for Building Management and Energy Efficiency" under grant number FO999923190.
Herbert Muehlburger: The position of this author is funded by the FFG project "The automated ontology generator" under grant number FO999901761.

Acknowledgements We thank the anonymous reviewers for their constructive and valuable feedback.

1 Introduction

Heating, ventilation, and air conditioning (HVAC) systems account for over 30% of global energy consumption, with heat pumps playing a critical role in modern energy-efficient buildings [3]. As cyber-physical systems (CPS), heat pumps generate high-dimensional sensor data that reflect complex operational states. Detecting faults in these systems is essential to ensure energy efficiency, equipment longevity, and user comfort.

Traditional fault detection and diagnosis (FDD) approaches rely on physics-based models or expert-defined rules [6], which require deep domain expertise and are difficult to scale. Supervised machine learning (ML) has emerged as a scalable alternative [6, 9], but it relies on large volumes of labeled fault data – often unavailable in real-world CPS deployments. To mitigate this issue, semi-supervised and weakly supervised methods [5, 16] have been explored, and regression-based field modeling approaches have also been proposed [8]. However, these still depend on handcrafted features or partial annotations.

Recent advances in large language models (LLMs) offer a new paradigm for fault detection. LLMs show strong generalization capabilities across diverse tasks, including structured reasoning and zero-shot classification. Prior work has explored their use in CPS contexts, such as prompt-based anomaly detection in battery systems [7], and time-series anomaly detection using structured inputs [1]. In addition, Xu et al. [15] proposed a benchmark called VisualTimeAnomaly that translates time series into visual representations, enabling multimodal LLMs to reason about sensor dynamics. However, LLM performance often lags behind deep learning models unless guided by carefully engineered prompts.

Meanwhile, foundation models for time series forecasting have shown that prompt structure and in-context reasoning play a key role in zero-shot generalization. TimesFM [4] demonstrates strong forecasting performance using a decoder-only architecture trained on time-series data, and TiRex [2] improves forecasting via context distillation and hybrid memory attention. These developments highlight the importance of architectural and representational choices in zero-shot time-series tasks.

Building on these insights, we ask: can LLMs detect faults in heat pump systems using only data from normal operation – without labeled fault data or expert rules?

To answer this, we introduce **GEMMA-FD**, a zero-shot, multimodal LLM framework for fault detection in CPS. We reformulate sensor-based fault detection as a structured prompting task: multivariate time series are converted into natural language descriptions enriched with statistical summaries, trend cues, and operating modes. We augment these textual prompts with visual inputs (e.g., heatmaps, line plots, histograms) to guide multimodal reasoning in GEMMA-3 [13], an open-weight, vision-and-language multimodal LLM released by Google DeepMind, in a purely zero-shot setting. No fine-tuning or labeled faults are required. Details on prompt construction and visual summarization are described in Section 3.2.

Our evaluation shows that GEMMA-3 tends to identify normal operating states correctly, but struggles to detect rare faults – reflecting the current limitations of zero-shot classification in highly imbalanced CPS settings. Nevertheless, the model captures CPS dynamics to a surprising extent and offers a reproducible, interpretable, and multilingual baseline for low-resource FDD.

Our main contributions are:

- We present GEMMA-FD, a zero-shot fault detection framework using GEMMA-3, which transforms sensor time series into structured natural language prompts without requiring labeled fault data.
- We introduce prompt designs that incorporate domain-specific interpretations, sensor correlations, and visual features (e.g., heatmaps, timeseries plots, histograms) to enhance LLM reasoning over complex system behavior.
- We benchmark GEMMA-3 against a supervised XGBoost baseline trained on the same prompts, highlighting the trade-offs between zero-shot generalization and supervised learning for rare fault detection.
- We release a fully reproducible pipeline for heat pump fault detection with open-weight LLMs, including data preprocessing, prompt generation, and evaluation code.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details our methodology and dataset. Section 4 presents and discusses the experimental results. Finally, Section 6 concludes the paper and outlines directions for future research.

2 Related Work

Fault Detection and Diagnosis (FDD) in cyber-physical systems (CPS), including HVAC and heat pump systems, has traditionally relied on model-based or data-driven methods. Classical approaches use physics-based models or expert-defined rules [6], which require significant domain knowledge and manual effort. As sensor data has become more abundant, supervised machine learning (ML) methods have emerged [6], although they rely on labeled fault data that are often costly and limited in diversity.

Several studies have benchmarked ML algorithms for heat pump FDD. Rahman et al. [9] compared XGBoost, Random Forest, SVM, and k-NN, identifying XGBoost as the most effective. Weigert et al. [14] utilized smart meter data for operational and anomaly classification. Shin and Cho [11] predicted the Coefficient of Performance (COP) to support FDD. Other work, such as by Sunal et al. [12], applied deep learning to fault detection in centrifugal pumps, offering insights transferable to HVAC systems. To reduce dependency on labeled data, semi-supervised and weakly supervised methods have been explored [5, 16]. Puttige et al. [8] demonstrated how regression and neural networks trained on field data can model heat pump behavior effectively.

Time Series and LLM-based Anomaly Detection. Recent advances in large language models (LLMs) have introduced new opportunities for time series anomaly detection. Alnegheimish et al. [1] explored whether general-purpose LLMs can act as zero-shot anomaly detectors and found that structured prompting is essential, though performance often trails deep learning baselines. Xu et al. [15] proposed AnomalyTransformer, a self-attention-based approach for modeling dependencies in time series anomalies. The same authors introduced VisualTimeAnomaly, a benchmark that translates time series into visual representations to evaluate multimodal LLM performance on TSAD tasks.

Foundation models tailored to time series have further advanced zero-shot reasoning. TimesFM [4] presents a decoder-only transformer trained on large-scale time series for forecasting tasks. TiRex [2] builds on this by improving in-context forecasting performance across short and long horizons via context distillation and hybrid memory attention. While both focus on forecasting, their findings inform how architectural and prompt design choices affect zero-shot generalization for time-dependent tasks.

LLMs in Fault Detection. Muehlburger et al. [7] demonstrated prompt-based anomaly detection in battery systems using open-weight LLMs, without requiring fine-tuning. In the HVAC domain, Hofer and Wotawa [6] showed that supervised learning informed by expert knowledge can yield high-performance FDD models for heat pumps.

Despite their generalization capacity, current LLMs often struggle with zero-shot classification of rare anomalies unless augmented with domain-specific cues, visual summaries, or few-shot examples. Our work builds on these insights by introducing a fully prompt-based, multimodal, zero-shot fault detection framework. We show that open-weight foundation models can partially encode CPS operational manifolds through structured prompts and visualizations – though further enhancement is needed for reliable minority-class classification.

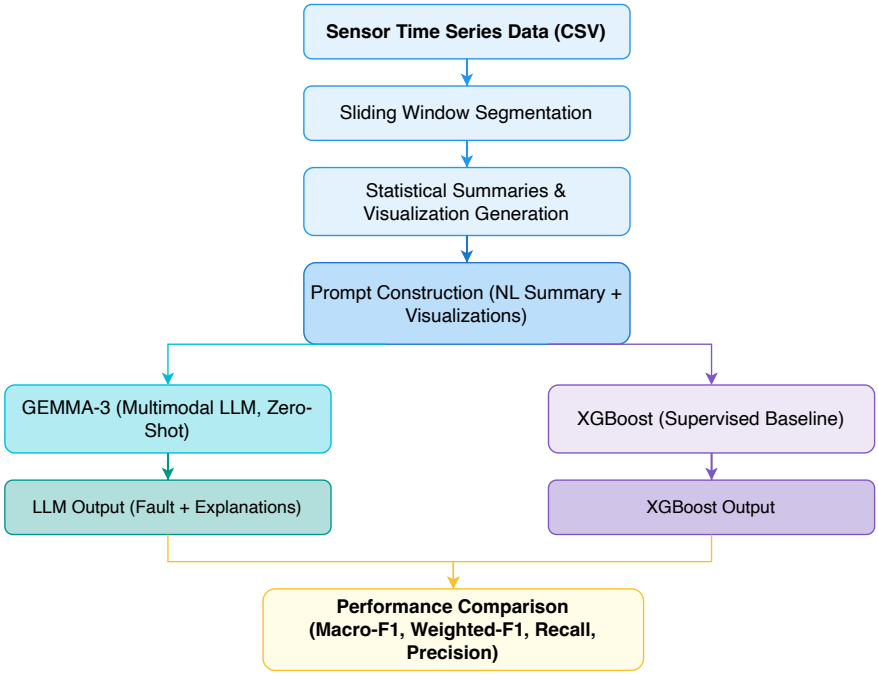


Figure 1 Overview of the GEMMA-FD pipeline. Sensor data is windowed and summarized with statistics and visualizations. Structured prompts are fed into GEMMA-3 for zero-shot inference, while the same text is used in a supervised XGBoost baseline. Outputs are compared using standard performance metrics.

To the best of our knowledge, this is the first application of an open-weight, zero-shot LLM-based method for multilingual, interpretable, and label-free fault detection in heat pump systems. We position GEMMA-FD as a diagnostic baseline, illuminating the strengths and limitations of current foundation models in real-world CPS deployments.

3 Methodology

We compare two contrasting diagnostic pipelines for fault detection in heat pump systems: a supervised baseline using XGBoost and a zero-shot multimodal baseline using GEMMA-3. Both pipelines consume the same structured prompt format but differ in knowledge acquisition: XGBoost is trained on labeled data, while GEMMA-3 infers labels in a zero-shot setting without any task-specific training.

- **Supervised baseline (XGBoost):** A gradient-boosted decision tree model trained on structured prompt-label pairs derived from sensor data.
- **Zero-shot baseline (GEMMA-3):** An open-weight multimodal large language model (LLM) that receives structured prompts and visual summaries as input and performs inference-only classification.

This setup allows us to assess the trade-off between supervised accuracy and zero-shot generalization in low-resource diagnostics. Figure 1 illustrates the overall fault detection pipeline (XGBoost had been trained before).

3.1 Dataset and Pipeline Overview

We use multivariate time series data logged from an Austrian heat pump system. Each row in the CSV logs represents a 1-minute snapshot of multiple synchronized sensor readings. The dataset contains operational data of the heat pump system (AnlageA_JKj), annotated into four categories: `betrieb_ok` (normal operation), `defrosting_issue`, `driver_temp_error`, and `overheating_control_issue`. The dataset contains in total 130,351 samples. For model training and evaluation we split the data into training (70%, 91,245 samples), validation (15%, 19,553), and test (15%, 19,553) sets using a stratified sampling strategy. We further follow two different data pre-processing paths:

- **XGBoost path:** Each row is treated as an individual sample for supervised training and testing.
- **GEMMA-3 path:** Data is segmented into overlapping 4-hour windows. Each window is converted into a textual prompt and three visualizations.

All samples are transformed into structured natural language prompts (see Section 3.2). For GEMMA-3, the prompt is enriched with composite images (heatmap, time series, and histogram) to support multimodal reasoning.

The GEMMA-FD pipeline processes multivariate time series data logged from a real-world heat pump system. Sensor readings are stored as CSV files, where each row represents a 1-minute interval of synchronized measurements across multiple physical variables (e.g., temperatures, flow rates, compressor speed). We implement two processing pathways:

- A **row-wise** path for supervised learning (XGBoost) that treats each minute independently.
- A **windowed** path for zero-shot prompting (GEMMA-3) that aggregates sensor readings over a 4-hour sliding window to provide temporal context.

As illustrated in Algorithm 1 each sample (row or window) is transformed into a **structured textual prompt**. For GEMMA-3, additional **multimodal visual summaries** (heatmap, time series, histogram) are also generated and jointly passed to the model.

Algorithm 1 GEMMA-FD Pipeline Overview.

Require: Multivariate time series $\mathbf{X}$; annotated row-wise labels (for supervised baseline)

if Supervised **then**
  for each row x_i in $\mathbf{X}$ **do**
    Construct structured prompt $P(x_i)$ from features
  end for
  Train XGBoost on TF–IDF representations of $P(x_i)$
else
  for each time window W_i in $\mathbf{X}$ **do**
    Extract statistical summaries and trends
    Generate time-series plot, heatmap, and histogram
    Construct structured natural language prompt $P(W_i)$
    Pass $P(W_i)$ and corresponding image to GEMMA-3 for zero-shot classification
  end for
end if
return Predicted class labels $\hat{y}$

3.2 Prompt Generation

To enable large language models (LLMs) to reason about system state, we convert raw heat pump sensor data into structured textual summaries, using a sliding-window approach. Each window corresponds to a fixed temporal segment (e.g., 4 hours) of multivariate time-series data, which is then transformed into a prompt as follows:

- **Sensor statistics**: For each relevant sensor (e.g., compressor speed, flow rate, inlet/outlet temperatures), we compute the mean, minimum, and maximum over the window. These are presented in natural language.
- **Trend detection**: For selected sensors, we detect increasing or decreasing trends using the first and last sample in the window.
- **Operating mode inference**: If available, the categorical "wp-app-state_38" field is used to infer the dominant operational mode (e.g., heating, standby, defrost).
- **Visual summary (multimodal input)**: A composite image is generated per window containing: (i) a z-score normalized heatmap, (ii) raw time-series plots, and (iii) sensor histograms. This image is passed to the LLM alongside the text.
- **Instructions for classification**: We append an instruction block asking the LLM to jointly analyze the image and text and classify the window into one of four predefined fault categories.

These design choices aim to emulate how human experts summarize system behavior and enable the LLM to leverage both symbolic and visual cues for anomaly detection. An example prompt is shown below:

Listing 1 Zero-shot user-prompt for GEMMA-3 with statistical summaries and classification instructions.

```
=== HEAT PUMP SENSOR WINDOW SUMMARY ===
Samples: 60 (2023-01-10 08:00 to 2023-01-10 12:00)
Most frequent operating mode: heating mode

Sensor statistics:
T2 Außen: mean=2.41, min=-1.00, max=5.20
wp-comp-speed: mean=2800.33, min=2600.00, max=3200.00
...

Trends:
Compressor speed increased over the window.

Instructions:
- The image below contains a Z-score normalized heatmap and time series plots for all sensors.
- Analyze the numeric data and the image.
- Look for abnormal periods, outliers, or persistent trends.
- Classify the window as one of: [betrieb_ok, defrosting_issue, driver_temp_error, overheating_control_issue].
- Output ONLY three function calls as specified. Reply in English and German.
```

Listing 1 presents a representative user prompt automatically generated using lightweight Python functions, while Listing 2 shows the corresponding system prompt provided to the model. This fully automated setup eliminates the need for manual annotation and enables scalable, LLM-based analysis of large, unlabeled datasets.

3.3 Model Configuration

We use a locally hosted, open-weight multimodal language model `gemma3:4b`, available through the Ollama framework. This variant supports both text and image input and is based on the GEMMA-3 architecture released by Google DeepMind in 2025. Table 1 summarizes key model characteristics.

Table 1 LLM configuration for windowed heat pump analysis.

Property	Value
Model	gemma3:4b (multimodal)
Parameters	4 billion
Weights	Open-source (Google DeepMind GEMMA-3)
Inference engine	Ollama v0.1.34
Modality	Multimodal (image + text)
Context window	8,192 tokens
Hardware	Apple MacBook Pro 14.2", M1 Pro, 10 Core CPU, 16 Core GPU
Average inference time	∼2 seconds per window
Batching	Single-sample (per window)
Prompt tuning	None (zero-shot only)

All predictions were performed using zero-shot prompting, without any fine-tuning or few-shot examples. Each prompt-image pair is sent independently to the model. The outputs are parsed to extract fault labels and free-text explanations.

3.4 Zero-Shot Fault Detection with GEMMA-3

We select `gemma3:4b`, an open-weight, multimodal LLM released by Google DeepMind [13] and served via the Ollama framework, to maximize reproducibility and local deployment feasibility. This 4B-parameter model supports both text and image inputs and is well-suited for vision-language inference tasks in constrained environments.

All experiments are conducted locally, with each sample consisting of a structured textual prompt and three visualizations as described in Section 3.2. The model receives a diagnostic system prompt specifying the output format; only the predicted label from the `report` function call is used for evaluation, parsed via regular expressions.

Evaluation is performed on a stratified sample of 300 test prompts. This setup reflects the label-scarce, zero-shot conditions of real-world fault detection. Our method is inspired in part by VisualTimeAnomaly [15], which highlights the value of multimodal visualizations for time series anomaly detection, and by Alnegheimish et al. [1], who emphasize the role of structured prompt engineering for LLM-based time series analysis.

To enrich the LLM's context, each inference window is described by a structured natural language prompt and three visualizations: a z-score normalized heatmap of all sensor signals (Figure 2a), a time series plot (Figure 2b), and histograms of value distributions for each sensor (Figure 2c). All visual inputs are directly derived from real-world heat pump sensor

Listing 2 System prompt provided to GEMMA-3 for multimodal zero-shot fault classification in heat pumps. The model outputs three Python function calls: `report()`, `diagnostics()`, and `diagnostics_de()`.

```
You are a diagnostic assistant for heat pump systems.
For each sample, you will receive:
- A plot (attached image) showing a 4-hour window of all sensor data (heatmap, time series, and histogram)
- A structured textual summary of that window

Your task:
- Use BOTH the plot and the textual summary to classify the system state during this period.
- Consider all short or persistent abnormalities, not just the overall trend.
- If you are unsure, explain your reasoning in the diagnostics output.
- All classes are equally likely.

Output format:
- Output exactly three Python function calls, in this order:
   1. report(fault_type: str)
   2. diagnostics(explanation: str) # (English)
   3. diagnostics_de(explanation: str) # (German)

Valid fault types:
- betrieb_ok: No abnormalities detected during the 24h period.
- defrosting_issue: Signs of persistent or recurring defrosting problems.
- driver_temp_error: Driver circuit temperature is out of expected range.
- overheating_control_issue: Evidence of overheating or failed thermal regulation.

Example output:
report('driver_temp_error')
diagnostics('Between 03:00 and 04:00, the driver temperature exceeded the normal range. The rest of the day appears normal.')
diagnostics_de('Zwischen 03:00 und 04:00 lag die Fahrertemperatur auß erhalb des Normalbereichs. Der Rest des Tages war unauffällig.')

IMPORTANT: Do not output any explanations or extra text outside of these three function calls. Do not include code blocks or define any functions.
```

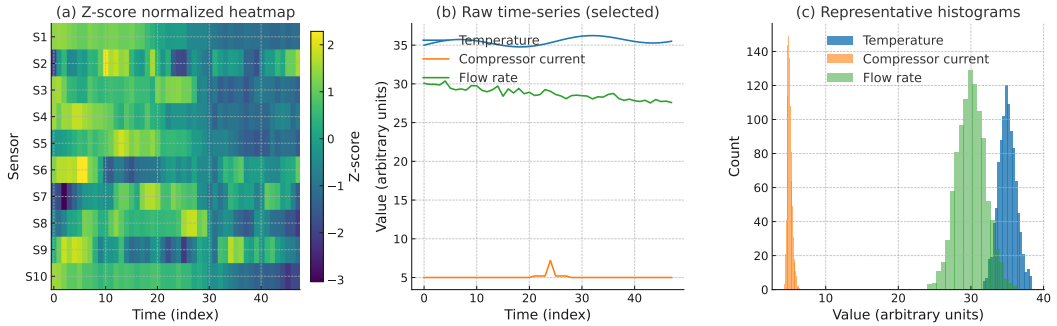


Figure 2 Visual encodings used as multimodal inputs to GEMMA-3. Each time window is represented as **(a)** a z-score normalized heatmap of sensor signals, **(b)** selected raw time-series plots, and **(c)** representative histograms of sensor value distributions. The values shown are exemplary and schematic for clarity; actual sensor values differ in scale and range but are visualized in the same encoding format. This schematic representation highlights the multimodal input structure (temporal, statistical, distributional) rather than domain-specific magnitudes.

data collected in Austria. The axis labels and variable names in the generated plots appear in German, as they reflect the naming conventions of the original control system and are passed to GEMMA-3 without translation to preserve semantic fidelity. These visuals capture temporal and statistical patterns, supporting robust zero-shot anomaly and fault detection.

For the zero-shot fault detection experiments, we evaluated GEMMA-3 using a stratified random sample of 300 test windows. To ensure class balance in the evaluation set, we performed stratified sampling across all annotated system states. Each sample consisted of a structured prompt and multimodal visualizations, as described previously.

The prediction workflow is as follows:

1. Stratify and sample $N = 300$ test windows from the full labeled dataset, preserving class proportions.
2. For each sampled window, generate a structured prompt and corresponding visualizations.
3. Run the `llm_predict` function, which queries GEMMA-3 via the Ollama inference framework with each prompt-image pair.
4. Collect and store all raw LLM outputs for further analysis and reproducibility.
5. Parse predicted fault labels from the structured LLM output using regular expressions.
6. Save all predictions and explanations to CSV and TXT files for downstream analysis.
7. Compute classification metrics (precision, recall, F1-score) and confusion matrix on the held-out samples.
8. Export results as LaTeX tables and plots for inclusion in the paper.

All LLM experiments are conducted with the model running locally to ensure full reproducibility and data privacy. We restrict the evaluation to 300 test samples per run to manage computational demands and facilitate interpretability of individual LLM predictions.

3.5 Supervised Baseline: XGBoost

For the supervised baseline, we train an XGBoost classifier using the full row-level dataset, where each row corresponds to a sensor snapshot and is labeled with one of the four classes: `betrieb_ok`, `defrosting_issue`, `driver_temp_error`, or `overheating_control_issue`. Individual CSV files for each class are parsed and transformed into prompt-label pairs using domain-informed prompt generation. After concatenating the class-specific data, we

split the combined dataset into training, validation, and test partitions. To address class imbalance, the training set is balanced using random oversampling. Each prompt is then vectorized using term frequency-inverse document frequency (TF–IDF) features, allowing up to 2,000 unigrams and bigrams.

The XGBoost classifier (`objective="multi:softprob"`) is trained on the balanced training data with hyperparameters set to 200 trees, a learning rate of 0.05, and a maximum tree depth of 6. After training, the model is evaluated on the held-out test partition, with classification metrics and confusion matrices generated to assess performance. This process is summarized in the following steps:

1. Load and concatenate row-level CSV data for each fault class.
2. Generate structured prompt-label pairs per sample.
3. Stratify into train/validation/test splits.
4. Apply random oversampling to balance the training set.
5. Vectorize prompts with TF–IDF.
6. Train the XGBoost classifier and evaluate on the test set.
7. Report class-wise precision, recall, F1-score, and confusion matrix.

This workflow describes our supervised baseline for comparison with zero-shot LLM-based fault detection.

3.6 Evaluation Metrics and Protocol

We report precision, recall, and F1-score for each class, as well as macro- and weighted averages to account for class imbalance and fault rarity, following Seliya et al. [10]. Confusion matrices visualize misclassifications.

Evaluation Protocols. For computational feasibility and qualitative inspection, GEMMA-3 was evaluated on 300 stratified, randomly-sampled windows from the test set, ensuring balanced class representation. XGBoost was evaluated on the full test set of 19,553 samples. Here are the details for the two settings:

- **Windowed Zero-Shot LLM (GEMMA-3):** We selected a stratified random sample of $N = 300$ test windows to ensure balanced class representation. Each sample was processed into a structured prompt and multimodal visualization. This sample size was chosen to balance computational feasibility and qualitative interpretability of individual LLM outputs.
- **Row-Wise and XGBoost (Full Test Set):** For the row-wise GEMMA-3 variant and the supervised XGBoost baseline, we evaluated on the complete held-out test set ($n \approx 19,553$ samples). No additional stratification was applied; all labeled samples were included for the XGBoost training.

For each protocol, predictions were parsed, performance metrics computed, and confusion matrices generated. This dual protocol quantifies both balanced and full-scale model performance for zero-shot and supervised approaches.

Metric definitions. Precision ($\frac{TP}{TP+FP}$) measures the proportion of correctly identified faults among flagged cases; recall ($\frac{TP}{TP+FN}$) measures the proportion of actual faults correctly detected; F1-score ($2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$) is their harmonic mean.

4 Results

This section presents a comparative analysis of fault detection in heat pump systems using three approaches: (1) GEMMA-3 in a zero-shot setting with windowed, multimodal prompts, (2) GEMMA-3 with row-wise, non-windowed prompts, and (3) a supervised XGBoost baseline. We evaluate all methods on the same structured dataset, reporting class-wise precision, recall, and F1-scores, as well as macro- and weighted averages to account for class imbalance. Quantitative results are complemented by qualitative analyses of LLM output.

4.1 Overall Performance

GEMMA-3, evaluated on a stratified random sample of 300 windowed test prompts, achieved a macro-F1 score of 0.19 and a weighted F1 of 0.39 (Table 2), indicating reliable detection of the dominant class but limited sensitivity to rare faults. XGBoost, in contrast, reached a macro-F1 of 0.69 and a weighted F1 of 0.83 on the full test set (19,553 samples; Table 4), demonstrating robust, balanced performance across all fault categories. These results underscore the advantage of supervised learning with labeled data for comprehensive fault detection, while highlighting the generalization challenge faced by zero-shot LLM-based approaches.

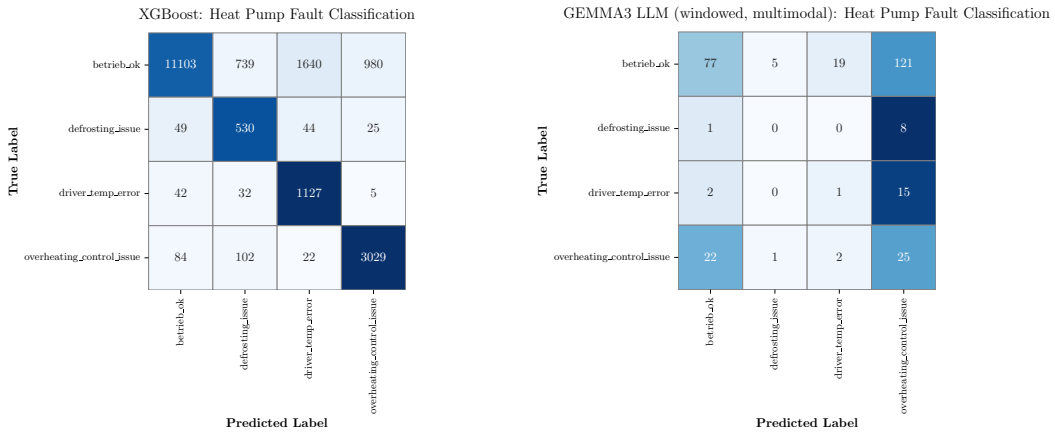
We observe that GEMMA-3 achieves its highest recall on the normal class (`betrieb_ok`), suggesting that it captures expected system behavior more reliably than anomalous regimes, consistent with prior findings on LLM robustness to typical inputs [1].

4.2 Class-Wise Performance Trends

A detailed examination of class-wise metrics reveals that GEMMA-3 achieves its highest recall for the majority class, `betrieb_ok`, but performs poorly on the minority fault classes. Specifically, GEMMA-3 attains a recall of 0.345 for `betrieb_ok`, while recall drops to zero for `defrosting_issue` and to 0.056 for `driver_temp_error`. The model achieves a moderate recall of 0.500 for `overheating_control_issue`, but with low precision. In contrast, the supervised XGBoost model shows high recall across all classes, including 0.818 for `defrosting_issue`, 0.934 for `driver_temp_error`, and 0.936 for `overheating_control_issue`. This demonstrates that XGBoost not only identifies the majority class but also maintains strong detection performance for rare faults.

4.3 Error Patterns and Limitations

Analysis of the confusion matrix for GEMMA-3 (Figure 3b) shows that the model predominantly assigns samples to the majority class, `betrieb_ok`, resulting in frequent misclassification of fault cases as normal operation. Both `defrosting_issue` and `driver_temp_error` are rarely, if ever, correctly identified, with most of these cases assigned to the normal class. For `overheating_control_issue`, GEMMA-3 correctly identifies some instances but also produces a high number of false positives. In contrast, the confusion matrix for XGBoost (Figure 3a) is dominated by high values along the diagonal, indicating correct classification for all classes and minimal confusion between normal and fault types.



(a) XGBoost on full test set.

(b) GEMMA-3 on 300 sampled prompts.

Figure 3 Confusion matrices comparing supervised and zero-shot diagnostic performance. **(a)** XGBoost achieves high accuracy across all fault classes with supervised training. **(b)** GEMMA-3, operating in a zero-shot setting, displays strong bias toward the majority class (`betrieb_ok`), with limited sensitivity to rare faults.

4.4 Performance Comparison

To further clarify the differences between approaches, we present a direct comparison of zero-shot LLM classification with GEMMA-3 and supervised learning with XGBoost. Both models are evaluated on identical input data and assessed using standard classification metrics. The following subsections provide detailed results for each method.

4.4.1 Zero-Shot Inference with Windowing

Table 2 shows the results for the zero-shot setting, where GEMMA-3 was evaluated on 300 randomly selected windows of test prompts. The model output was parsed to extract the predicted fault class. Figure 3b shows the confusion matrix for this evaluation.

Table 2 Classification performance of GEMMA-3 for zero-shot fault detection in heat pump systems using structured prompts and windowed multimodal data. The model shows strong recall for normal operation but limited accuracy for rare fault classes, underscoring the challenges of zero-shot detection without labeled data.

Class	Precision	Recall	F1	Support
betrieb_ok	0.755	0.345	0.474	223
defrosting_issue	0.000	0.000	0.000	9
driver_temp_error	0.045	0.056	0.050	18
overheating_control_issue	0.148	0.500	0.228	50
micro avg	0.344	0.343	0.344	300
macro avg	0.237	0.225	0.188	300
weighted avg	0.589	0.343	0.393	300

The model assigned the majority of samples to the `betrieb_ok` class, resulting in a recall of 0.345 for this category. For the rare fault classes, recall was near zero for `defrosting_issue` and 0.056 for `driver_temp_error`, while recall for `overheating_control_issue` reached 0.500 but with low precision. The macro-F1 score was 0.188, and the weighted-F1 was 0.393, indicating overall limited effectiveness in rare fault detection.

Table 3 Classification performance of GEMMA-3 for zero-shot fault detection in heat pump systems using row-wise, non-windowed prompts. Compared to the windowed setting, the model shows higher recall for some fault classes but still struggles to detect `overheating_control_issue`, indicating limitations in distinguishing complex fault patterns without temporal context.

Class	Precision	Recall	F1	Support
betrieb_ok	0.743	0.706	0.724	14,051
defrosting_issue	0.046	0.251	0.078	630
driver_temp_error	0.159	0.299	0.207	1,170
overheating_control_issue	0.000	0.000	0.000	0
micro avg	0.549	0.658	0.598	15,851
macro avg	0.237	0.314	0.252	15,851
weighted avg	0.672	0.658	0.660	15,851

4.4.2 Row-Wise Zero-Shot Inference (No Windowing)

Table 3 reports classification performance for GEMMA-3 in a row-wise zero-shot setting, where each time step is treated independently without temporal aggregation. The model was evaluated on the entire test set ($n = 15{,}851$) using non-windowed prompts.

The highest F1-score was achieved for the normal operating state `betrieb_ok` (F1 = 0.724), with precision and recall of 0.743 and 0.706, respectively. This indicates that the model performs reasonably well for detecting normal operation. Fault classes, however, exhibit significantly lower scores. For example, `defrosting_issue` achieved a recall of 0.251 but extremely low precision (0.046), indicating frequent false positives. Similarly, `driver_temp_error` reached a recall of 0.299 with low precision (0.159), suggesting limited fault discrimination ability.

Notably, the model completely failed to detect any instances of `overheating_control_issue`, as indicated by zero precision, recall, and F1 score likely due to lack of context in single-timestep prompts. This aligns with the observation that more complex temporal patterns are difficult to infer without windowed context.

The macro-averaged F1-score of 0.252 and the weighted average of 0.660 reflect the model's bias toward majority classes. Although the row-wise approach scales more easily to large datasets, it fails to capture time-dependent anomaly patterns essential for robust fault diagnosis. These findings highlight inherent challenges in employing multimodal LLMs directly for fault detection tasks, especially for minority classes. Potential strategies to enhance performance include targeted prompt-tuning, data augmentation techniques, or incorporating ensemble methods to improve robustness and accuracy.

4.4.3 Supervised XGBoost Baseline

The supervised XGBoost model was trained and evaluated on the same dataset as the LLM, using class-weighted loss to address class imbalance. It achieved, as shown in Figure 4, a macro-average F1 score of 0.693 and a weighted-average F1 score of 0.827. Recall exceeded 0.81 for all fault classes, with the highest values observed for `overheating_control_issue` (0.936) and `driver_temp_error` (0.934).

Table 4 Classification performance of XGBoost trained on supervised fault labels for heat pump fault detection. XGBoost demonstrates high precision and recall across all classes, including rare faults, highlighting the effectiveness of supervised learning when labeled data are available.

Class	Precision	Recall	F1	Support
betrieb_ok	0.984	0.768	0.863	14,462
defrosting_issue	0.378	0.818	0.517	648
driver_temp_error	0.398	0.934	0.558	1,206
overheating_control_issue	0.750	0.936	0.833	3,237
macro avg	0.627	0.864	0.693	19,553
weighted avg	0.889	0.807	0.827	19,553

These results demonstrate that supervised, class-aware training yields strong diagnostic performance, even when the input consists of structured natural language prompts identical to those used by the LLM.

4.5 Error Analysis

A detailed examination of the GEMMA-3 results (Tables 2 and 3) reveal three dominant error patterns:

- **Majority class bias:** GEMMA-3 disproportionately predicts the `betrieb_ok` (normal operation) class, even for samples labeled with faults. This leads to very low recall for rare fault types such as `defrosting_issue` (0.000) and `driver_temp_error` (0.056), reflecting a strong bias toward the majority class. This behavior is clearly visible in the confusion matrix (Figure 3b).
- **Limited fault-specific detection:** Fault samples are rarely classified with the correct fault type, even when statistical summaries contain values outside expected ranges. This suggests that GEMMA-3, operating under zero-shot conditions, struggles to associate specific numerical patterns with fault semantics.
- **Overprediction of certain faults:** For `overheating_control_issue`, the model exhibits moderate recall (0.500) but low precision (0.148). This pattern suggests that the model often predicts this fault incorrectly (false positives), indicating a tendency to overtrigger this class rather than true ambiguity at the decision boundary.

In contrast, the XGBoost baseline achieves high recall and precision for all classes, including minority faults (Table 4). Its confusion matrix (Figure 3a) is dominated by correct predictions, with minimal misclassification between fault types. This contrast highlights the challenge of zero-shot LLM-based fault detection – namely, that generalization from unlabeled normal data is insufficient for robust fault classification in complex cyber-physical systems.

These findings underscore the need for future work on:

- Enhancing prompt design with explicit fault indicators and temporal patterns,
- Applying few-shot learning to expose models to representative fault cases,
- Leveraging visual modality inputs (e.g., plots) more effectively for multimodal reasoning.

4.6 Comparison and Insights

A direct comparison between GEMMA-3 and XGBoost highlights the fundamental trade-off between label-free generalization and supervised precision in heat pump fault detection.

- **XGBoost consistently outperforms GEMMA-3**, particularly on rare faults such as `defrosting_issue` and `driver_temp_error`, achieving high F1-scores across all classes – even when trained on natural language prompts originally designed for LLMs. This underscores the effectiveness of supervised learning when annotated data is available, regardless of input format.
- **GEMMA-3 enables fully label-free fault detection**, operating in a zero-shot regime using structured prompts and visualizations. However, it struggles with class imbalance and underperforms on minority fault types, limiting its current practical utility.
- **LLM-based diagnostics remain promising**, especially for low-resource or rapidly evolving environments. Hybrid strategies that integrate LLMs with few-shot exemplars, domain knowledge, or statistical post-processing may help overcome zero-shot limitations.

These results emphasize that while foundation models enable flexible deployment without labeled data, they are not yet reliable for safety-critical diagnostics. Improving multimodal prompt design and combining statistical learning with LLM reasoning are key directions for future research.

4.7 Future Directions

Out current framework already incorporates multimodal prompts – combining structured text with visual representations such as heatmaps, time series, and histograms. Future work could focus on leveraging these features more effectively. Enhancing prompt engineering with explicit fault indicators and temporal patterns, as well as introducing few-shot learning with annotated fault cases, may help address the low recall for rare fault types. In addition, hybrid approaches that combine LLM-based reasoning with supervised or statistical models could further improve robustness and accuracy. These directions aim to bridge the remaining gap between zero-shot generalization and reliable fault detection in complex, real-world cyber-physical systems.

In summary, supervised learning with XGBoost yields balanced detection across all fault types, while GEMMA-3's zero-shot, prompt-based classification is largely limited to majority-class detection and fails to generalize to rare faults. These findings highlight the fundamental trade-off between the flexibility of zero-shot LLMs and the reliability of supervised models, motivating the improvements discussed in the next section.

5 Discussion and Limitations

Our results reveal a core trade-off in fault detection for cyber-physical systems: supervised models like XGBoost achieve high accuracy across all classes (macro-F1 = 0.69) by leveraging labeled data, while GEMMA-3 enables interpretable, multilingual zero-shot diagnostics (macro-F1 = 0.24) without requiring any labeled faults. However, GEMMA-3 suffers from majority-class bias and limited recall for rare anomalies.

This performance gap underscores the limitations of zero-shot prompting when fault-specific cues are weak or absent. Nonetheless, GEMMA-3 delivers natural language explanations and can be deployed in settings where labeled data is scarce or unavailable, making it a viable foundation for low-label diagnostic pipelines.

To improve zero-shot fault detection, we propose three extensions:

1. **Enhance prompt engineering** with explicit temporal and fault-indicative features.
2. **Introduce few-shot prompts** to provide representative fault exemplars during inference.
3. **Incorporate retrieval-augmented generation (RAG)** using domain-specific corpora (e.g., technical manuals, field logs) to support root cause explanation.

In this study, we excluded RAG and fine-tuning to isolate the effects of prompt structure and visual context. This allows us to establish GEMMA-FD as a reproducible zero-shot baseline and benchmark for future hybrid approaches.

We also acknowledge the limitation of evaluating on a proprietary dataset. Future work will expand to public HVAC datasets (e.g., UCI SECOM, AHU) and incorporate unsupervised methods (e.g., Isolation Forest, LSTM Autoencoders) to assess generalizability and complement LLM reasoning.

Our findings suggest that structured prompting with visual encodings provides a viable entry point for fault detection in low-label regimes – but bridging the gap to robust deployment will require hybrid approaches that combine LLMs with symbolic models, supervision, or retrieval.

6 Conclusion and Outlook

We introduced **GEMMA-FD**, a zero-shot, prompt-based framework for fault detection in heat pumps using open-weight, multimodal large language models. By converting sensor data windows into structured text and visual inputs, our approach enables interpretable fault classification without labeled anomalies or expert rules.

While GEMMA-3 shows promise as a flexible, label-free diagnostic tool, it underperforms supervised methods like XGBoost in precision and recall, particularly for rare fault types. This illustrates the core trade-off between generalization and accuracy in CPS fault detection.

Looking forward, we see strong potential in hybrid approaches that combine the interpretability and accessibility of LLMs with the precision of supervised models. Enhancements such as few-shot adaptation, prompt ensembling, and retrieval-augmented generation can bridge current gaps and improve fault isolation and explanation in real-world deployments.

Ethical Statement

This research does not involve any ethical concerns or conflicts of interest.

References

1 Sarah Alnegheimish, Linh Nguyen, Laure Berti-Equille, and Kalyan Veeramachaneni. Can Large Language Models be Anomaly Detectors for Time Series? In *2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, October 2024. doi:10.1109/DSAA61799.2024.10722786.

2 Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex: Zero-Shot Forecasting Across Long and Short Horizons with Enhanced In-Context Learning, May 2025. doi:10.48550/arXiv.2505.23719.

3 Elaheh Bazdar, Fuzhan Nasiri, and Fariborz Haghighat. Optimal planning and configuration of adiabatic-compressed air energy storage for urban buildings application: Techno-economic and environmental assessment. *Journal of Energy Storage*, 76:109720, January 2024. doi:10.1016/j.est.2023.109720.

4 Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 10148–10167, Vienna, Austria, July 2024. JMLR.org.

5 Hongliang Fei, Younghun Kim, Sambit Sahu, Milind Naphade, Sanjay K. Mamidipalli, and John Hutchinson. Heat pump detection from coarse grained smart meter data with positive and unlabeled learning. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1330–1338, New York, NY, USA, August 2013. Association for Computing Machinery. doi:10.1145/2487575.2488203.

6 Birgit Hofer and Franz Wotawa. Detecting Soft Faults in Heat Pumps (Short Paper). *OASIcs, Volume 125, DX 2024*, 125:22:1–22:10, 2024. doi:10.4230/OASICS.DX.2024.22.

7 Herbert Muehlburger and Franz Wotawa. FaultLines - Evaluating the Efficacy of Open-Source Large Language Models for Fault Detection in Cyber-Physical Systems*. In *2024 IEEE International Conference on Artificial Intelligence Testing (AITest)*, pages 47–54, July 2024. doi:10.1109/AITest62860.2024.00014.

8 Anjan Rao Puttige, Staffan Andersson, Ronny Östin, and Thomas Olofsson. Application of Regression and ANN Models for Heat Pumps with Field Measurements. *Energies*, 14(6):1750, January 2021. doi:10.3390/en14061750.

9 Md Mahbubur Rahman, Reza Malekian, and Vilhelm Akerstroem. Fault Detection On Heat Pump Operational Data Using Machine Learning Algorithms. In *2024 11th International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*, pages 204–211, September 2024. doi:10.1109/IOTSMS62296.2024.10710259.

10 Naeem Seliya, Taghi M. Khoshgoftaar, and Jason Van Hulse. A Study on the Relationships of Classifier Performance Metrics. In *2009 21st IEEE International Conference on Tools with Artificial Intelligence*, pages 59–66, November 2009. doi:10.1109/ICTAI.2009.25.

11 Ji-Hyun Shin and Young-Hum Cho. Machine-Learning-Based Coefficient of Performance Prediction Model for Heat Pump Systems. *Applied Sciences*, 12(1):362, January 2022. doi:10.3390/app12010362.

12 Cem Ekin Sunal, Vladimir Dyo, and Vladan Velisavljevic. Review of Machine Learning Based Fault Detection for Centrifugal Pump Induction Motors. *IEEE Access*, 10:71344–71355, 2022. doi:10.1109/ACCESS.2022.3187718.

13 Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, and Ramona Merhej. Gemma 3 Technical Report, March 2025. doi:10.48550/arXiv.2503.19786.

14 Andreas Weigert, Konstantin Hopf, Nicolai Weinig, and Thorsten Staake. Detection of heat pumps from smart meter and open data. *Energy Informatics*, 3(1):21, October 2020. doi:10.1186/s42162-020-00124-6.

15 Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy, June 2022. doi:10.48550/arXiv.2110.02642.

16 Wei Yin, Guo-qing Wang, Wan-sheng Miao, Min Zhang, and Wei-guo Zhang. Semi-supervised learning of decision making for parts faults to system-level failures diagnosis in avionics system. In *2012 IEEE/AIAA 31st Digital Avionics Systems Conference (DASC)*, pages 7C4–1–7C4–14, October 2012. doi:10.1109/DASC.2012.6382418.