

H T
W I
G N

Hochschule Konstanz
Department of Computer Science

Submitted by
Samuel Tim
Student Number 307636

samuel.tim200@yahoo.de

B

C



Bachelor Thesis

Energy Anomaly Detection with Machine Learning

S



Konstanz, 31st December 2025

Bachelor Thesis

Energy Anomaly Detection with Machine Learning

by

Samuel Tim

in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Science

in Applied Computer Science

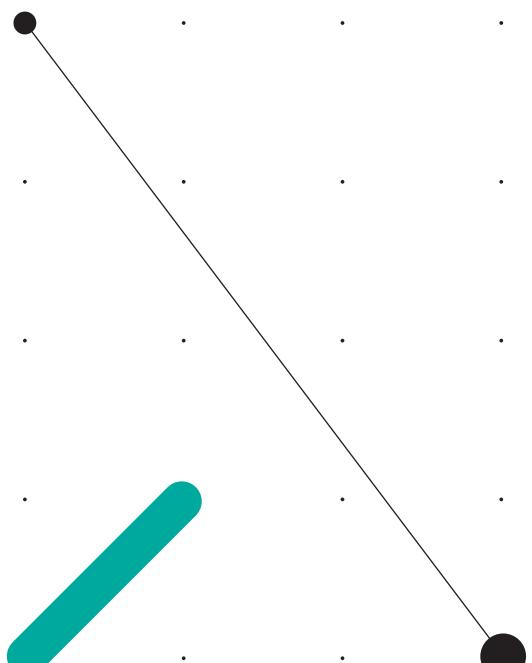
at the Hochschule Konstanz University of Applied Sciences,

Student Number: 307636

Date of Submission: 31st December 2025

Supervisor: **Prof. Dr. Marko Boger**

Second Examiner: **Dipl.-Inf. Björn Erb**



An electronic version of this thesis is available at <http://repository.tudelft.nl/>.

Abstract

Buildings account for approximately 30% of global final energy consumption, while empirical studies estimate that between 4% and 18% of building energy use is attributable to anomalies such as technical faults, control and scheduling errors, and behaviour-induced energy misuse, which result in avoidable operational inefficiencies.

This thesis presents an integrated methodology for contextual anomaly detection in multivariate, non-stationary building-energy time series, enabling financial-impact estimation and automated root-cause attribution. The approach is fully implemented within an existing IoT building-management platform and includes a production-ready frontend for anomaly visualization and operational analysis.

To address structural limitations of existing anomaly-detection benchmarks for building-energy data, a dedicated evaluation dataset was constructed using the BOPTEST simulation environment, comprising a clean baseline and systematically injected multivariate, context-dependent anomaly scenarios. Multiple detection methods, including statistical, deep-learning, and foundation-model-based approaches, were evaluated on this benchmark.

The results indicate that stochastic prediction models with probabilistic output distributions are more suitable than deterministic point predictors for modelling multimodal building-energy behaviour and identifying contextual anomalies. Chronos-2 enables the practical application of time-series foundation models to multivariate energy telemetry without per-asset training, while mixture-density modelling was identified as a promising architectural direction for future research. The findings establish a methodological basis for a universal energy foundation model supporting zero-shot anomaly detection and standardized baseline comparison in accordance with IPMVP.

Contents

1	Introduction	1
1.1	Motivation and Economic Context	1
1.2	Digitalization of Buildings and Data Explosion	1
1.3	Why Current Building Automation Systems Fail	2
1.4	Emergence of Foundation Models for Time Series	3
1.5	System Context and Industrial Relevance	4
1.6	Problem Scope and Research Contributions	4
2	Foundations	7
2.1	Characteristics of Building Energy Data	7
2.1.1	Multivariate Structure	7
2.1.2	Causal Chain of Energy Consumption	8
2.1.3	Temporal Dependence and Persistence	10
2.1.4	Seasonality and Periodicity	10
2.1.5	Statistical Distribution and Non-Stationarity	10
2.1.6	Data Acquisition and Semantic Structure	11
2.1.7	Data Continuity and Transmission Artifacts	12
2.1.8	Synthesis of Data Characteristics	12
2.2	Foundations of Anomaly Detection	13
2.2.1	Dimensionality and Normality Regimes	13
2.2.2	Structural Classes of Anomalies	13
2.2.3	Multiplicity of Occurrence	15
2.2.4	Research Scope	15
2.3	Methodological Approaches to Anomaly Detection	15
2.3.1	Anomaly Scores	15
2.3.2	Learning Paradigms	15
2.3.3	Families of Detection Methods	17
2.4	Benchmarking Foundations	18
2.4.1	Binary Labeling and Ground Truth	18
2.4.2	The Confusion Matrix: Four Possible Outcomes	18
2.4.3	Measuring Success: Precision, Recall, and the F1 Score	19
2.5	Synthesis of Foundations	20

3 Related Work	21
3.1 Reliability and Benchmarking: The TSB-AD Framework	21
3.1.1 Systemic Flaws and Metric Reliability	21
3.1.2 Benchmark Evaluation and Model Hierarchy	22
3.1.3 Implications for Multivariate Context Point Anomalies	23
3.2 Comparative Analysis of Deep Learning and Foundation Models in Energy Systems	24
3.2.1 Deep Generative Models and the Advantage of Reconstruction	24
3.2.2 The Emergence of Time-Series Foundation Models	24
3.2.3 Synthesis: Prediction-Based Stochastic Modeling	25
3.3 TODO: Mixture Density Networks for Stochastic Modeling	25
3.4 TODO: Root Cause Analysis for Anomaly Detection	26
3.5 Identification of Research Gaps	26
3.5.1 Representation Gap in Multivariate Context Point Anomalies	26
3.5.2 Temporal Baseline and Generalization Gap	26
3.5.3 Architectural Gap in Stochastic Mixture Modeling	27
3.5.4 Diagnostic and Economic Functional Gap	27
3.6 Synthesis of Research Objectives	27
4 Methodology	29
4.1 System Context: The Eliona IoT Platform	29
4.1.1 Modular System Architecture	29
4.1.2 Asset Modeling and Hierarchical Ontology	30
4.2 System Requirements Specification	30
4.2.1 Functional Requirements	31
4.2.2 Operational and Data Integrity Requirements	32
4.3 Proposed High-Level Architecture	32
4.3.1 Component Interaction and Data Flow	32
4.3.2 Integration into the Azure Ecosystem	33
4.4 Proof of Concept (PoC): Stochastic Feasibility	34
4.4.1 MDN-Based Stochastic Prediction	34
4.4.2 Diagnostic Integration: SHAP and LLM	34
4.5 Critique of Sequential Forecasting for Anomaly Detection	34
4.5.1 Synthetic Experimental Setup	35
4.5.2 Failure Mode 1: Error Propagation and Instability	35
4.5.3 Failure Mode 2: Rapid Adaptation and the PA-F1 Illusion	36
4.5.4 Mitigation Strategies	37

4.6	Statistical Limitations of Point and Gaussian Predictions	38
4.6.1	The Failure of Mean Squared Error Minimization	39
4.6.2	The Gaussian Distribution Paradox	40
4.6.3	Solution: Mixture Density Networks	40
4.7	Benchmark Data Generation and Composition	43
4.7.1	Simulation Environment and Baseline Construction	43
4.7.2	Feature Selection and Control Layer Logic	43
4.8	Experimental Data Segmentation and Anomaly Injection	44
4.8.1	Segmentation Strategy	44
4.8.2	Anomaly Taxonomy and Labeling	45
5	Implementation	47
5.1	Integrated System Architecture and Technology Stack	47
5.1.1	Deployment and Cluster Integration	47
5.1.2	Data Orchestration and Persistence	48
5.2	Python Analytics Endpoint: Chronos-2 Integration	48
5.2.1	Predictive Logic and Model Hosting	49
5.2.2	Managed Online Endpoints in Azure ML	49
5.3	Scala Microservice: Multi-Tenant Orchestration	49
5.3.1	Multi-Tenant Lifecycle Management	50
5.4	Data Acquisition and Processing Pipeline	50
5.4.1	Attribute Filtering and Hierarchical Selection	51
5.4.2	Contextual Enrichment: Site-Localized Weather	51
5.4.3	Data Cleaning and Gap-Resilience Logic	51
5.4.4	Reactive Data Fetching	52
5.5	Stochastic Inference and Anomaly Quantification	52
5.5.1	Feature-Driven Prediction Strategy	52
5.5.2	Batch Processing and Quantile Requests	53
5.5.3	Detection Logic and Financial Quantification	53
5.6	Hierarchical Root Cause Analysis (RCA)	54
5.6.1	Diagnostic Attribution	54
5.6.2	Localization and Weather Context	54
5.7	Temporal Collapse and Persistence	55
5.8	AI Synthesis and Recommended Actions	55
5.8.1	Scenario A: Behavioral Fault (Lighting)	55
5.8.2	Scenario B: Technical Fault or Misuse (Plug Loads)	55

5.9	Tenant-Specific Configuration and Parameterization	56
5.10	Frontend Visualization and User Interaction	57
5.10.1	The Anomalies Table Interface	57
5.10.2	User Feedback and Status Management	57
5.10.3	Integrated Anomaly Analytics and Visualization	58
5.10.4	Anomaly Detail View and Operational Synthesis	61
5.10.5	Anomaly Statistics and Macro-Level Reporting	61
5.10.6	Asset-Specific Anomaly Integration	64
6	Discussion and Future Work	65
6.1	Critical Reflection on System Design	65
6.2	Data Integrity and User-Centric Baseline Selection	66
6.3	Future Architecture: The Universal Energy Feature Forecaster	66
6.3.1	In-Context Zero-Shot Modelling	66
6.3.2	Probabilistic Anomaly Scoring	67
6.4	Reflections on Energy Anomaly Benchmarking	67
7	Conclusion	69
	References	71

Glossary

anomaly observation or pattern that deviates significantly from a defined notion of normality. [vi](#), [9](#)

benchmark standardized dataset and evaluation protocol used to compare the performance of different anomaly detection methods. [vi](#)

confusion matrix tabular summary of prediction results that counts true positives, true negatives, false positives, and false negatives. [vi](#), [22](#)

ground truth reference labels that indicate for each observation whether it is considered normal or anomalous, used as a standard when evaluating detection performance. [vi](#), [21](#), [22](#)

mislabeling inconsistent assignment of anomaly labels to similar or identical patterns, which distorts evaluation by inflating false-negative rates. [vi](#), [25](#)

precision for anomaly detection, the proportion of predicted anomalous points or segments that are actually anomalous (true positives divided by all positive predictions). [vi](#), [22](#), [23](#)

recall for anomaly detection, the proportion of truly anomalous points or segments that are correctly detected (true positives divided by all actual anomalies). [vi](#), [22](#), [23](#)

run-to-failure bias systematic placement of anomalies at the end of a time series, which favors models that exploit positional cues rather than genuine signal deviations. [vi](#), [25](#)

unrealistic anomaly ratio an artificially high proportion of anomalous observations in a dataset compared to real-world systems, which can lead to over-optimistic performance estimates. [vi](#), [26](#)

Acronyms

BL base load. [9](#)

CNN convolutional neural network. [26–28](#)

DL deep learning. [28](#)

F1 F1 score. [22, 23](#)

FM foundation model. [28](#)

FN false negative. [22](#)

FP false positive. [22](#)

GAN generative adversarial network. [28](#)

HVAC heating, ventilation and air conditioning. [8, 9, 13, 15, 17](#)

IoT Internet of Things. [28](#)

LSTM long short-term memory network. [28](#)

ML machine learning. [27](#)

OmniAnomaly stochastic recurrent neural network model OmniAnomaly. [26, 28](#)

PA-F1 Point-Adjustment F1 score. [26](#)

RE reconstruction error. [28](#)

RP reconstruction probability. [28](#)

SOTA state of the art. [5](#)

Sub-PCA subspace principal component analysis. [26](#)

TN true negative. [22](#)

TP true positive. [22](#)

TSAD time series anomaly detection. [5](#), [14](#), [25](#), [28](#)

TSB-AD Time Series Benchmark for Anomaly Detection. [26](#), [27](#)

TSFM time-series foundation model. [28](#)

VAE Variational Autoencoder. [28](#)

VUS-PR Volume Under the Surface–Precision Recall. [26](#), [27](#)

1

Introduction

1.1. Motivation and Economic Context

Buildings account for approximately 30% of global final energy consumption and more than 50% of global electricity consumption [Alš24]. Empirical studies indicate that avoidable operational anomalies—encompassing technical faults, suboptimal control strategies, and persistent behavioural misuse—account for between 4% and 18% of building energy use [Rot+04]. These inefficiencies frequently remain undetected because conventional threshold-based monitoring systems are not triggered.

Non-technical losses represent a quantifiable economic burden. Electricity theft results in annual losses exceeding 6 billion USD in the United States alone [MM09]. Furthermore, reports from the World Bank indicate that in some developing countries up to 50% of distributed electricity is lost due to theft [Ant09]. Such patterns of energy misuse constitute an economically relevant class of anomalies. While modern **Building Automation Systems (BAS)** are capable of detecting deviations from nominal operation, they typically neither quantify the associated financial impact nor provide systematic root-cause attribution, thereby limiting their operational and economic usefulness.

1.2. Digitalization of Buildings and Data Explosion

The implementation of **Advanced Metering Infrastructure (AMI)**, which combines smart meters with communication networks, is expanding globally. In the United States, smart

meters had been deployed for approximately 77% of households and businesses by 2022, with the installed base projected to grow to about 134 million devices in 2024 and 142 million in 2026 [Edi24]. The increasing integration of digital infrastructure and sub-metering in modern building environments generates vast repositories of high-frequency telemetry. This abundance of data provides a unique opportunity for the application of advanced artificial-intelligence techniques that thrive on large-scale, high-resolution multivariate data to identify previously undetectable deviations.

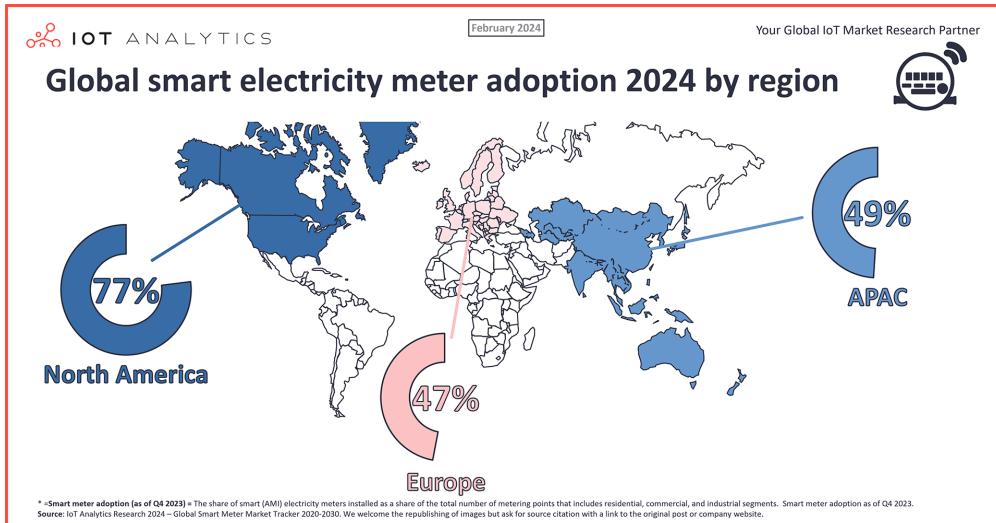


Figure 1.1: Global smart electricity meter adoption by region in 2024, illustrating the varying levels of AMI penetration across markets [IoT24].

1.3. Why Current Building Automation Systems Fail

Most deployed Building Automation Systems (BAS) rely on static rule-based logic and univariate statistical thresholds applied to individual sensor streams. Such approaches are structurally incapable of capturing the multivariate, context-dependent nature of building-energy behaviour and are therefore unable to distinguish between legitimate operational regime changes and true anomalous states.

More recent data-driven and machine-learning-based detection methods exhibit fundamental limitations. Many approaches operate on deterministic point predictions that fail to represent the stochastic, multimodal, and non-stationary characteristics of building-energy time series. As a result, these models impose context-agnostic deviation boundaries that treat identical absolute residuals as equally anomalous across fundamentally different operational regimes, leading to structurally incorrect anomaly semantics.

Sequential forecasting-based detectors further suffer from two critical failure modes when applied to sustained anomalies: (i) error propagation, where anomalous values corrupt the sliding input window and destabilize future predictions, and (ii) rapid baseline adaptation, where models quickly absorb anomalous states as normal operation, causing long-duration anomalies to disappear from the anomaly score signal :contentReference[oaicite:1]index=1. These effects undermine both detection reliability and financial loss quantification.

Furthermore, the majority of published anomaly-detection benchmarks rely on inadequately annotated datasets, implicit anomaly assumptions, and global point-anomaly definitions that can be captured by trivial threshold rules but fail to represent contextual and multivariate operational anomalies. These limitations significantly reduce the transferability of reported performance to real-world building operation.

Finally, existing systems rarely provide automated root-cause attribution or translate detected deviations into quantifiable financial impact, thereby limiting their practical value for operational decision-making and maintenance prioritization.

1.4. Emergence of Foundation Models for Time Series

The emergence of [time-series foundation model \(TSFM\)](#), such as Chronos-2 [Ans+25], marks a new paradigm in building-energy analytics. These models are designed to process multivariate, non-stationary, and stochastic data and provide probabilistic output distributions instead of single-value predictions. This enables the construction of normative operational bands that capture multimodal building behaviour and allow contextual deviations to be distinguished from normal variability.

A key advantage of foundation models is their zero-shot generalization capability. In contrast to asset-specific forecasting models, TSFMs do not require per-meter training or frequent retraining. Multivariate building telemetry can be provided directly as contextual input, while optional fine-tuning can be performed jointly across entire building portfolios. This makes foundation models particularly well suited to the inherently non-stationary nature of building-energy data and enables scalable deployment across large building estates.

1.5. System Context and Industrial Relevance

This research is conducted in the context of the Eliona IoT Building Management Platform (see Section 4.1), a production-grade multi-tenant system deployed in commercial and industrial building portfolios worldwide. Eliona integrates heterogeneous building automation systems, smart meters, and environmental sensors into a unified telemetry and analytics layer.

The anomaly-detection framework developed in this thesis is not a laboratory prototype, but a fully integrated subsystem within Eliona's operational architecture. It processes live building telemetry, performs stochastic anomaly detection, quantifies financial impact, localizes probable root causes, and exposes actionable insights through a production-ready frontend used by facility managers and energy operators.

This real-world deployment context defines both the functional requirements and the architectural constraints of the proposed methodology, including scalability, robustness to missing data, non-stationary baselines, explainability, and economic interpretability.

1.6. Problem Scope and Research Contributions

This thesis addresses the problem of detecting, economically quantifying, and diagnostically localizing contextual anomalies in multivariate building-energy time series within large-scale, non-stationary operational environments.

In contrast to traditional threshold-based and deterministic forecasting approaches, this work formulates anomaly detection as a stochastic, multivariate, context-dependent modeling problem. The objective is the design and implementation of an integrated, production-ready anomaly intelligence system that not only detects deviations, but also explains their technical origin, quantifies their economic impact, and derives operationally meaningful recommendations.

The proposed framework models expected building-energy behaviour as a multivariate, multimodal mixture distribution, allowing deviations to be evaluated probabilistically rather than against static thresholds. This formulation explicitly accounts for non-stationary baselines caused by seasonal shifts, occupancy changes, and long-term system drift, preventing legitimate regime transitions from being misclassified as anomalies. Methodologically, the work focuses on detecting [Multivariate Context Point Anomaly \(MCPA\)](#) and translating them into financially interpretable metrics.

The primary contributions of this thesis are:

- A formalization of building-energy anomaly detection as multivariate contextual point anomaly detection under multimodality and non-stationarity.
- An empirical and theoretical critique of deterministic forecasting-based anomaly detectors and their structural failure modes.
- A stochastic detection framework based on probabilistic normative bands derived from foundation models.
- A financially interpretable quantification layer that transforms deviations into monetary loss estimates.
- A hierarchical root-cause attribution pipeline grounded in building ontologies and causal dependencies.
- A domain-specific multivariate benchmark dataset generated via [Building Optimization Performance Test Framework \(BOPTEST\)](#) with labeled contextual fault scenarios (see Section 4.7).
- A fully integrated, scalable, multi-tenant implementation deployed within a production IoT building platform.
- A methodological foundation for the development of a universal energy foundation model supporting zero-shot anomaly detection and standardized baseline comparison in accordance with the [International Performance Measurement and Verification Protocol \(IPMVP\)](#).

This work assumes that the historical baseline used for model context represents nominal building operation. The framework is therefore designed to detect deviations emerging after baseline establishment and does not aim to retroactively identify faults that were already persistently present in historical reference data. Furthermore, the scope is limited to aggregated building-energy telemetry and does not target high-frequency electrical fault detection, equipment-level vibration analysis, or cybersecurity intrusion detection.

2

Foundations

This chapter establishes the theoretical and methodological foundations required for a comprehensive understanding of the subsequent research. The analysis begins with an examination of the technical characteristics and physical composition of building energy data to define the operational parameters. Subsequently, the fundamental concepts of anomalies are explored, encompassing the classification of specific types and their manifestation in temporal data. This positioning allows for the integration of the current use case into the broader field of [time series anomaly detection \(TSAD\)](#). Finally, Chapter 3 traces the technical evolution of detection methodologies and provides a systematic review of the current [state of the art \(SOTA\)](#).

2.1. Characteristics of Building Energy Data

Building-energy telemetry constitutes a multivariate, multimodal, and non-stationary stochastic process governed by physical, behavioural, and technical drivers. Effective anomaly detection therefore requires formal consideration of the structural and statistical properties of these data.

2.1.1. Multivariate Structure

Building energy data is inherently multivariate and interdependent. In addition to aggregate meter readings, relevant variables include environmental conditions, occupancy,

and subsystem states. Cross-variable dependencies are fundamental: changes in environmental drivers induce correlated changes in technical system loads. Consequently, anomaly detection must operate on multivariate joint behaviour rather than on isolated univariate series.

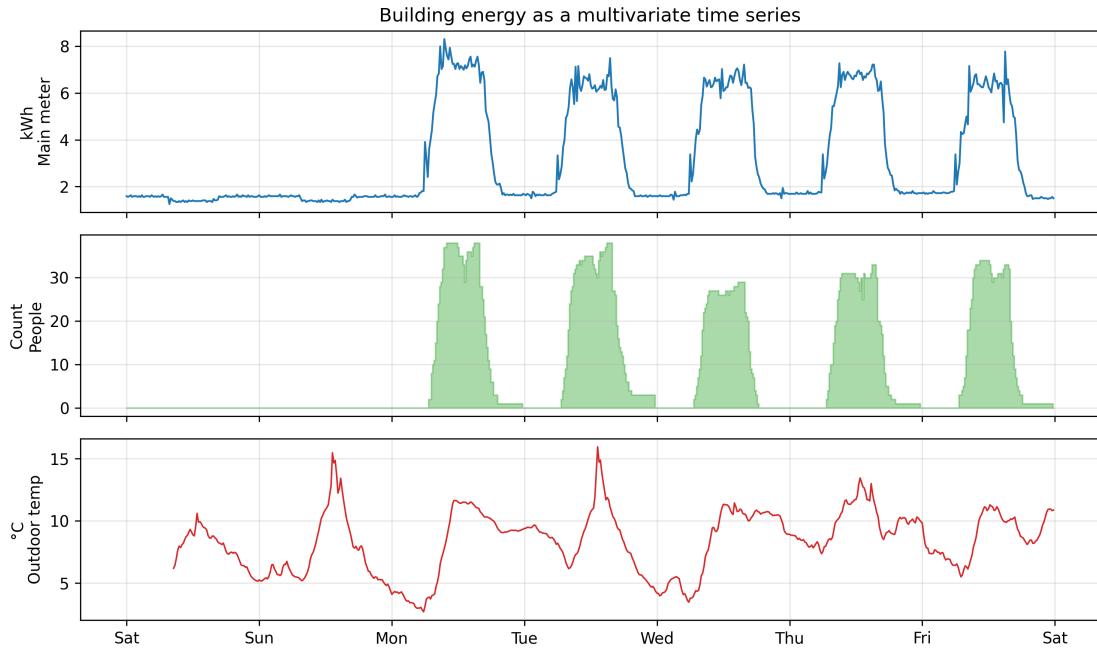


Figure 2.1: Representative multivariate time series showing the main meter load together with occupancy (people count) and outdoor temperature. The plot illustrates how multiple interdependent variables evolve jointly over time.

2.1.2. Causal Chain of Energy Consumption

Energy consumption emerges from a causal chain spanning demand generation, control logic, and mechanical execution. Environmental and occupancy conditions generate service demand; controllers translate demand into actuation commands; mechanical subsystems execute these commands, producing measurable energy use. Deviations observed at aggregate meters therefore frequently originate from faults located in upstream sensing or control layers.

Understanding the causal chain, as illustrated in Figure 2.2, is a prerequisite for localizing anomalous behavior within building systems. A deviation observed in the building's main meter often originates from a fault located in a preceding stage of the technical hierarchy, such as a sensor error or a logic failure in the control layer.

For instance, a malfunctioning temperature sensor reporting an erroneous heat spike triggers a cascade of responses. The control layer interprets this false data as

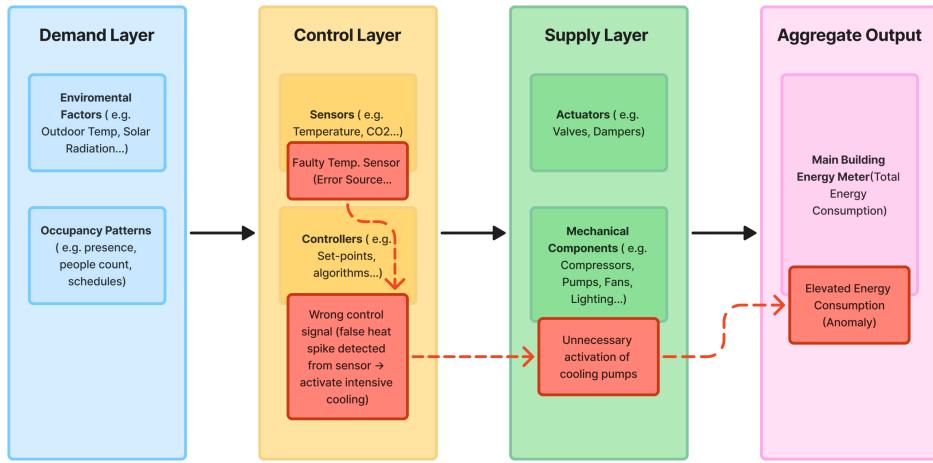


Figure 2.2: Causal chain of building energy consumption from demand over control to supply layer.

a thermal requirement and initiates a cooling command to counteract the perceived heat. This signal causes the supply layer to activate mechanical components, such as cooling pumps and compressors. These devices consume electrical energy to satisfy the requested cooling load. Consequently, the aggregate output layer, represented by the building's main energy meter, records a significant increase in consumption. In this scenario, the measured energy spike is not a result of an actual physical need but acts as a symptom of a failure located deeper in the technical hierarchy.

HVAC and Environmental Drivers

HVAC systems dominate building energy demand. Thermal gradients, solar radiation, humidity, and scheduling logic determine cooling and heating loads. Suboptimal control strategies, scheduling conflicts, and mechanical degradation induce baseline drift and excessive consumption, generating anomalies that are often operationally normal yet energetically inefficient.

Occupancy and Internal Loads

Human activity introduces stochastic variability through lighting, appliance use, and thermal gains. Behavioural interventions can decouple consumption from environmental drivers, while IT infrastructure introduces discrete operational regimes. These effects contribute to multimodality and regime-dependent energy patterns.

Structural Moderators and Data Integrity

Building envelope characteristics and thermal inertia modulate system response dynamics. Interdependencies between subsystems propagate anomalies across services. Digital measurement infrastructure introduces non-physical artifacts, including missing values and aggregation spikes, which must be distinguished from physical faults during preprocessing.

2.1.3. Temporal Dependence and Persistence

Building-energy telemetry exhibits strong temporal autocorrelation caused by thermal inertia, operational ramp-up dynamics, and persistent high-load device states. Consequently, short-term system behaviour is highly predictable under nominal operation, while slow-developing faults and sustained inefficiencies may remain concealed within otherwise smooth trajectories.

This persistence simultaneously stabilizes short-term forecasting and undermines detection of long-duration anomalies, particularly when sequential models rapidly absorb anomalous regimes into their predictive baseline.

2.1.4. Seasonality and Periodicity

Building-energy consumption follows pronounced daily, weekly, and seasonal periodicities driven by occupancy cycles, control schedules, and climatic seasons. These regime-dependent patterns form a repetitive operational fingerprint.

Anomaly detection must therefore distinguish contextual violations of expected periodic regimes (e.g., weekday-level consumption during weekends) from absolute deviations.

2.1.5. Statistical Distribution and Non-Stationarity

Empirical building-energy distributions deviate substantially from unimodal Gaussian assumptions and exhibit multimodal mixture structures with heavy tails due to discrete operational regimes and heterogeneous subsystem interactions. Deterministic point estimates are therefore insufficient to represent normative behaviour.

Sparse coverage of extreme weather and rare operational states introduces causal ambiguity and increases false anomaly rates for previously unobserved but physically valid conditions. Furthermore, building-energy telemetry is non-stationary; long-term baseline drift caused by seasonal transitions, equipment degradation, and persistent

occupancy changes continuously shifts normative distributions, necessitating probabilistic, context-aware modeling.

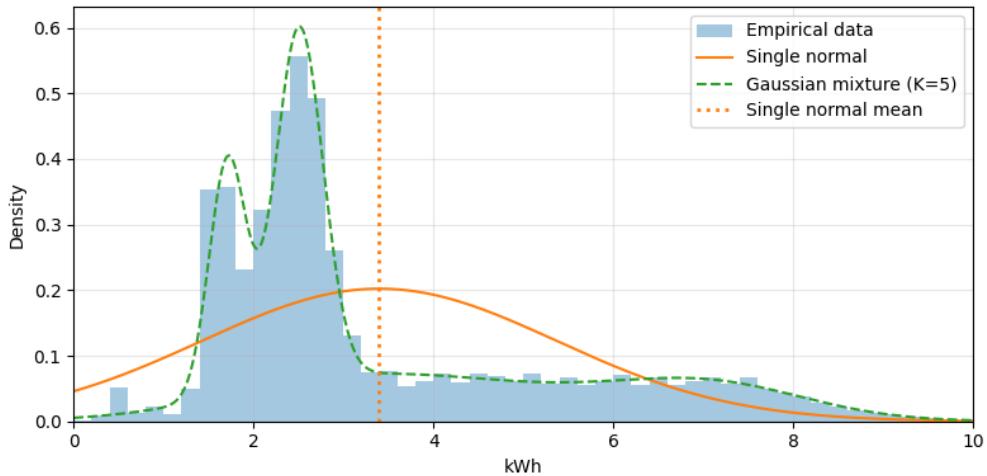


Figure 2.3: Empirical distribution of the building's main meter (15-minute kWh values, histogram) with an overlaid single normal distribution and a Gaussian mixture model with five components, illustrating the mismatch between a unimodal Gaussian model and the multimodal, heavy-tailed structure of real building energy data.

Figure 2.3 illustrates this effect on real data: the measured main-meter consumption concentrates in several dense regions at lower loads and exhibits a pronounced right tail. A single normal distribution smooths over these structures and underestimates tail probabilities, whereas the fitted Gaussian mixture adapts to the multiple modes and better traces the empirical density.

2.1.6. Data Acquisition and Semantic Structure

The transformation of physical energy consumption into digital telemetry follows a multi-stage acquisition pipeline that converts electrical quantities into structured multivariate time series suitable for algorithmic analysis.

Ontological Modeling: To ensure interpretability and causal localization, the telemetry is mapped to a semantic ontology that encodes the physical and logical relationships between meters, subsystems, and devices [Eli25c]. This ontological layer enables detected anomalies to be localized within the technical hierarchy rather than remaining aggregated deviations at the main meter level.

Standardized Units: Raw meter readings are converted into standardized physical units (e.g., kWh) to ensure consistency across heterogeneous hardware and communication interfaces.

2.1.7. Data Continuity and Transmission Artifacts

The integrity of telemetry streams depends on the stability of the communication infrastructure. Network-level distortions introduce non-physical artifacts that must be distinguished from actual building faults.

Transmission Gaps: Communication failures produce missing values that interrupt temporal continuity and require correction during preprocessing.

Aggregation Spikes: Buffered data retransmission following outages may produce virtual load spikes, reflecting delayed reporting rather than physical surges in energy demand.

2.1.8. Synthesis of Data Characteristics

Building-energy telemetry constitutes a multivariate, multimodal, and non-stationary stochastic process governed by a causal chain spanning demand, control, and mechanical execution. Anomalies observed at aggregate meters therefore represent manifestations of upstream technical and behavioural faults.

These data exhibit strong temporal autocorrelation and persistence, regime-dependent seasonality, multimodal mixture distributions, and digital transmission artifacts introduced by the acquisition pipeline. Consequently, effective anomaly detection must satisfy the following formal requirements:

1. Probabilistic multivariate modelling

Expected energy behaviour must be modelled probabilistically as a multivariate mixture distribution rather than as a deterministic point estimate, so that multimodal regimes and cross-variable dependencies are represented explicitly.

2. Robustness to temporal dependence and persistence

Detection must remain robust under strong autocorrelation, long-duration regime persistence, and seasonal pattern shifts, ensuring that sustained faults are not absorbed into the learned baseline.

3. Separation of physical faults and digital artifacts

Digital measurement artifacts—such as transmission gaps and aggregation spikes—must be distinguished from physical system anomalies through explicit preprocessing and data-quality handling.

Taken together, these characteristics and requirements formally define the modelling assumptions and architectural constraints that guide the anomaly detection methodologies developed in this thesis.

2.2. Foundations of Anomaly Detection

Anomaly detection aims to identify observations or patterns that deviate from an implicit notion of normality. In time-series anomaly detection (TSAD), deviations are defined relative to temporal structure, persistence, and regime-dependent behaviour rather than isolated numerical values. The taxonomy adopted in this work follows the benchmark framework proposed by Paparrizos et al. [Pap+22].

2.2.1. Dimensionality and Normality Regimes

The complexity of anomaly detection is governed by the dimensionality of the time series and the number of normative operational regimes.

Dimensionality: Univariate time series describe a single system variable, whereas multivariate time series jointly model multiple interdependent variables. The dataset analyzed in this work is multivariate, combining aggregate energy consumption with environmental and occupancy drivers to resolve causal ambiguity.

Normality Regimes: Building-energy telemetry operates under multiple normative regimes driven by seasonal, operational, and occupancy-dependent contexts. Consequently, “normal” behaviour is regime-specific rather than globally invariant.

The analyzed data is therefore classified as multivariate with multi-mode normality, necessitating detection models that adapt to shifting baselines and cross-variable dependencies.

2.2.2. Structural Classes of Anomalies

Point Anomalies: Individual observations that deviate from expected behaviour.

Global Point Anomalies: Deviations outside the global historical range.

Contextual Point Anomalies: Deviations from regime-dependent normative behaviour defined by temporal or exogenous context.

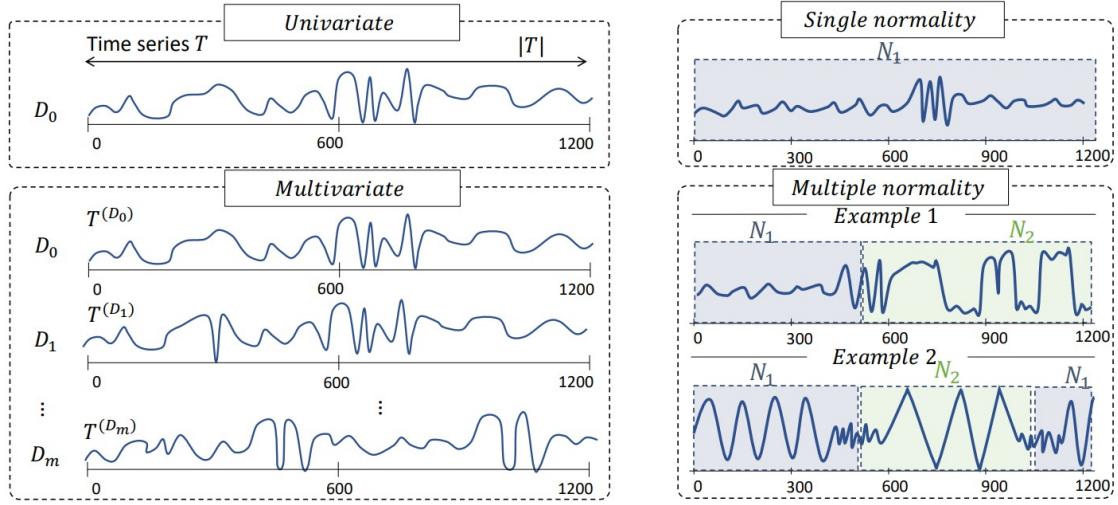


Figure 2.4: Schematic illustration of time series types along two axes—dimensionality (univariate vs. multivariate) and normality regimes (single-mode vs. multi-mode). Adapted from Boniol et al.'s tutorial on new trends in time series anomaly detection [Boniol2023NewTrends].

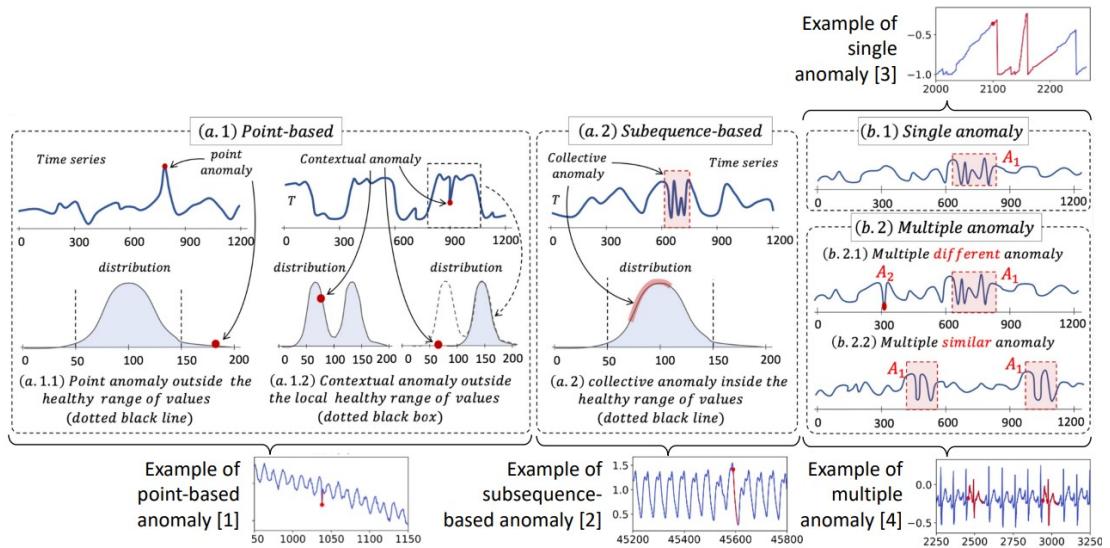


Figure 2.5: Taxonomy of time series anomalies along structural and multiplicity dimensions, distinguishing global and contextual point anomalies, subsequence-based anomalies, and their occurrence as single, multiple different, or multiple similar events. Adapted from Boniol et al.'s tutorial on new trends in time series anomaly detection [Boniol2023NewTrends].

Subsequence Anomalies: Deviations manifested through abnormal temporal patterns.

2.2.3. Multiplicity of Occurrence

Single Anomalies: Isolated anomalous events.

Multiple Similar Anomalies: Recurrent manifestations of the same anomaly pattern.

Multiple Different Anomalies: Co-occurring anomalies of heterogeneous types.

2.2.4. Research Scope

This work focuses on contextual point anomalies in multivariate building-energy telemetry. High-frequency subsequence anomalies manifest as contextual point deviations in aggregated energy series and are therefore addressed implicitly.

A central methodological challenge is the detection of multiple similar anomalies under non-stationarity, as recurrent inefficiencies may be absorbed into learned baselines. The proposed framework explicitly addresses this normality-drift problem through probabilistic modeling and controlled baseline adaptation.

2.3. Methodological Approaches to Anomaly Detection

Anomaly detection transforms raw time-series telemetry into actionable information by assigning each observation a degree of abnormality and mapping it to a binary decision boundary.

2.3.1. Anomaly Scores

Most detection algorithms output an anomaly score s_i per timestamp, which quantifies deviation from learned normative behaviour. Binary alerts are obtained by thresholding this score, yielding a time series of nominal and anomalous states.

2.3.2. Learning Paradigms

The applicability of detection methods is governed by the availability of labelled data:

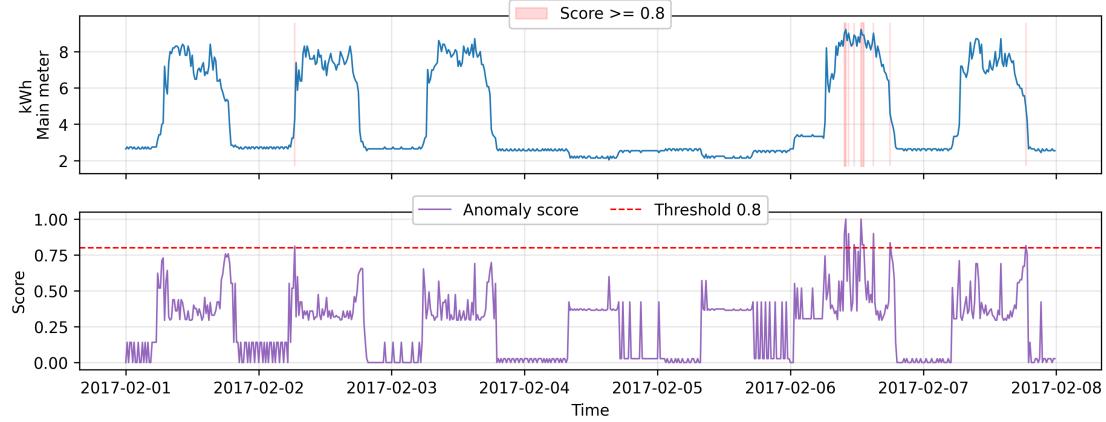


Figure 2.6: Example of an anomaly score $s_i \in [0, 1]$ aligned with the underlying time series. A threshold of 0.8 separates normal points from those flagged as anomalous.

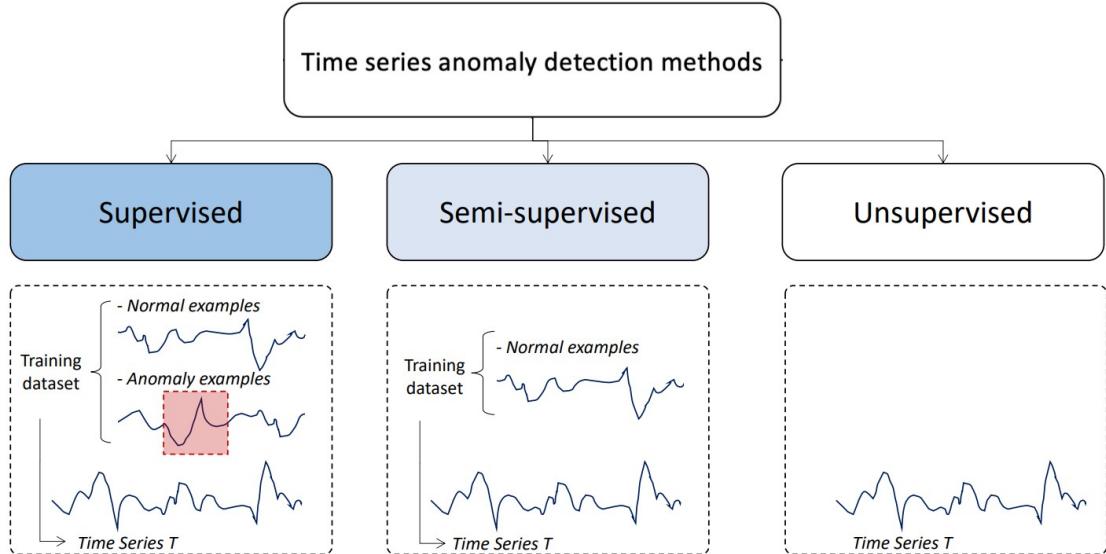


Figure 2.7: Schematic overview of supervised, semi-supervised, and unsupervised learning paradigms for anomaly detection, adapted from Boniol et al.'s tutorial on new trends in time series anomaly detection [[Boniol2023NewTrends](#)].

Supervised: Requires explicit labels for both normal and anomalous states; rarely feasible in building operations.

Semi-Supervised: Learns normative behaviour from assumed healthy historical data; commonly used in building-energy monitoring.

Unsupervised: Operates without labelled baselines; typically applied during system commissioning or cold-start phases.

2.3.3. Families of Detection Methods

Anomaly detection approaches are grouped into three methodological families:

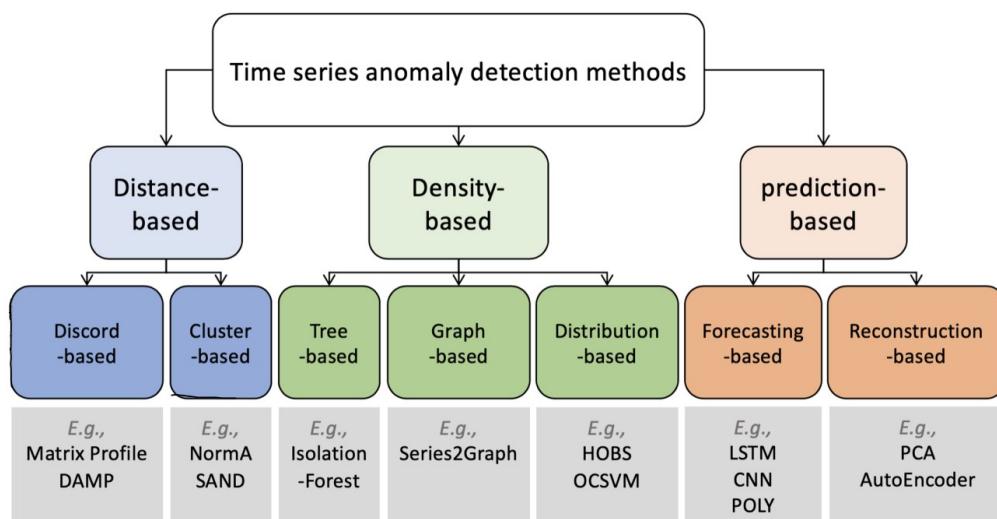


Figure 2.8: Hierarchical taxonomy of anomaly detection methods grouped into distance-based, density-based, and prediction-based families, adapted from Boniol et al.'s tutorial on new trends in time series anomaly detection [Boniol2023NewTrends].

Distance-Based: Identify anomalous subsequences by pattern dissimilarity.

Density-Based: Detect low-probability observations in learned feature distributions.

Prediction-Based: Identify deviations via residuals between predicted and observed values, including forecasting- and reconstruction-based variants.

This thesis focuses on prediction-based approaches, as they are the only methodological family that provides an explicit expected-value baseline, enabling deviations to

be quantified in physical units and directly translated into financial impact. This property is essential for contextual building-energy anomaly detection and economic loss estimation.

2.4. Benchmarking Foundations

To determine how well an anomaly detection system works, it must be evaluated against a dataset where the correct outcomes are already known. This process is referred to as benchmarking and relies on comparing the model's decisions with a reference set of labels, the [ground truth](#), to quantify detection performance.

2.4.1. Binary Labeling and Ground Truth

The foundation of any benchmark is a labeled dataset in which each observation is assigned a binary label indicating whether it is considered normal or anomalous. In this context, a label of 0 denotes normal operation and a label of 1 denotes an anomaly, such as a fault or unusual event. When an algorithm analyzes this dataset, it produces its own sequence of binary decisions. Benchmarking assesses how closely these model-generated labels align with the original [ground truth](#) labels.

2.4.2. The Confusion Matrix: Four Possible Outcomes

When the model's predictions are compared to the [ground truth](#), each observation falls into one of four categories. These outcomes are summarized in a [confusion matrix](#), which serves as a scorecard for the detection system:

True Positive ([true positive \(TP\)](#)): The model correctly identifies an anomaly; the ground truth label is 1 and the model predicts 1.

True Negative ([true negative \(TN\)](#)): The model correctly identifies normal operation; the ground truth label is 0 and the model predicts 0.

False Positive ([false positive \(FP\)](#)): The model raises a false alarm; it predicts an anomaly (1) while the ground truth label is normal (0).

False Negative ([false negative \(FN\)](#)): The model misses an anomaly; the ground truth label is 1 but the model predicts normal operation (0).

These four counts form the quantitative basis for most evaluation metrics used in anomaly detection benchmarks.

2.4.3. Measuring Success: Precision, Recall, and the F1 Score

The entries of the [confusion matrix](#) are used to derive summary metrics that characterize different aspects of a model's performance. Three central measures in anomaly detection are [precision](#), [recall](#), and the [F1 score \(F1\)](#) score.

[precision](#) quantifies how reliable the alarms are. It answers the question: when the model flags an anomaly, how often is this decision correct? Formally, precision is defined as the ratio of correctly detected anomalies to all observations that the model classified as anomalous,

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (2.1)$$

A high precision value indicates that the model rarely raises false alarms.

[recall](#) quantifies how many anomalies are successfully detected. It answers the question: of all anomalies that actually occurred, how many did the model identify? Recall is defined as

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (2.2)$$

A high recall value indicates that the model is sensitive to anomalous behavior and misses few faults.

In practice, there is often a trade-off between [precision](#) and [recall](#). A model that labels almost every point as anomalous may achieve high recall but very low precision, whereas an overly conservative model may exhibit the opposite behavior. The [F1](#) score provides a single scalar summary by combining precision and recall through their harmonic mean,

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (2.3)$$

A high [F1](#) score indicates that the model achieves a balanced compromise: it detects a large fraction of true anomalies while keeping the number of false alarms at a manageable level. In many benchmarking studies, this score is used as the primary indicator for ranking competing detection methods.

2.5. Synthesis of Foundations

The examination of building energy data in Section 2.1 and the formal taxonomies of anomaly detection in Section 2.2 reveal a highly specialized operational environment. Because building energy consumption is an aggregate signal driven by a complex causal chain (see Subsection 2.1.2), it is fundamentally characterized as a multivariate time series (see Subsection 2.1.1). This multidimensionality necessitates a focus on contextual point anomalies (see Subsection ??), as abnormality is defined primarily relative to the prevailing environmental and operational state rather than to a single global range. Furthermore, the temporal aggregation of energy data into 15-minute or hourly intervals (see Subsection 2.1.5) effectively transforms high-frequency rhythmic faults into detectable point-based deviations in the aggregated series.

The implementation of detection strategies is constrained by several domain-specific factors. Human-driven consumption and manual behavioral interventions introduce a stochastic element to the time series (see Subsection 2.1.2), where these probabilistic variations deviate from standard Gaussian models and require detection frameworks capable of distinguishing between random noise and technical faults (see Subsection 2.1.5). Simultaneously, the physical degradation of mechanical components causes a gradual increase in the power required for the same service output (see Subsection 2.1.2), ensuring that the definition of normality is non-stationary and undergoes baseline drift over long temporal scales.

Within this context, semi-supervised and unsupervised learning paradigms (see Subsection ??) introduce a fundamental technical contradiction when historical data is utilized in the absence of expert-annotated fault histories. If persistent or repetitive anomalies were already present during the training period, the model incorrectly incorporates these deviations into its learned definition of normality. This risk of normality drift, particularly in the presence of multiple similar anomalies (see Subsection ??), leads to suppressed alerts and remains a central constraint for the subsequent methodological selection.

The establishment of these data-driven requirements and theoretical classifications lays the foundation for a systematic evaluation of the current state of the art in Chapter 3, where concrete algorithmic choices are assessed against the constraints of building energy data.

3

Related Work

The identification of building energy anomalies requires an algorithmic framework that aligns with the structural and physical constraints established in Chapter 2. This chapter examines the current State of the Art (SOTA)—defined as the most advanced level of development in a technical field—to identify the most suitable detection methodology for the multivariate and contextual requirements of the research.

Existing research is evaluated through a systematic analysis of established benchmarks. The subsequent sections study these benchmarks to determine which methodologies demonstrate quantifiable results in resolving the complexities of energy consumption signals.

3.1. Reliability and Benchmarking: The TSB-AD Framework

The selection of an appropriate detection methodology is constrained by systemic issues within the existing research landscape. Liu and Paparrizos [LP24] identify these issues as the “elephant in the room,” demonstrating that apparent progress in TSAD is often an artifact of flawed evaluation practices rather than algorithmic superiority.

3.1.1. Systemic Flaws and Metric Reliability

Historical results are often compromised by three documented data-level flaws. First, **mislabeling** leads to artificially high false-negative rates. Second, a prevalent **run-to-**

failure bias rewards models that simply prioritize temporal position. Finally, unrealistic anomaly ratios fail to reflect the rarity of faults in physical systems.

The “illusion of progress” is further attributed to point-wise metrics like Point-Adjustment F1 score (PA-F1), which facilitates a significant overestimation of model performance by rewarding a detection if even a single point within an anomalous segment is identified. To ensure accuracy, this research adopts Volume Under the Surface–Precision Recall (VUS-PR), established by Liu and Paparrizos [LP24] as the robust standard for providing threshold-independent evaluation resistant to temporal lags and noisy scoring.

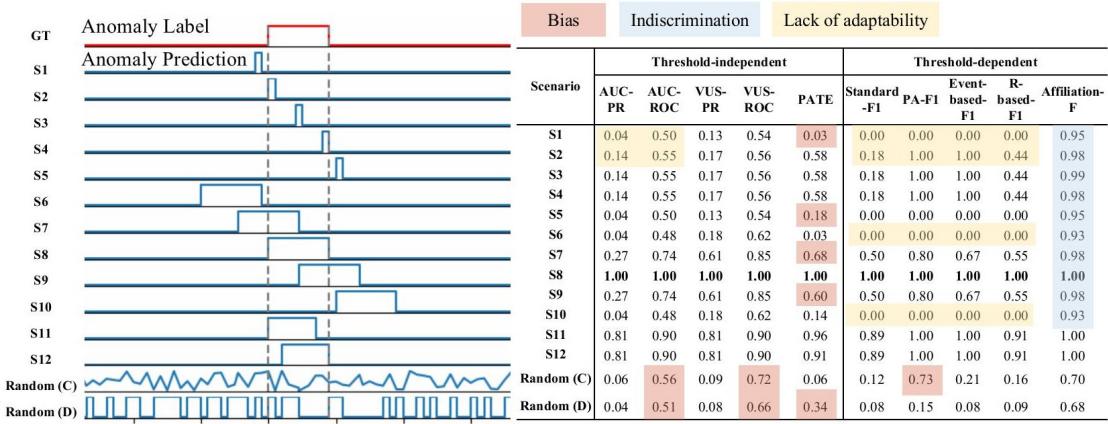


Figure 3.1: Reliability analysis of evaluation measures across different anomaly prediction scenarios. The red segment at the top represents the ground truth anomaly label, followed by various prediction signals (S1–S12 and random). The adjacent table indicates the resulting scores for threshold-independent and threshold-dependent metrics. Adapted from Liu and Paparrizos [LP24].

3.1.2. Benchmark Evaluation and Model Hierarchy

Evaluation across 1 070 curated time series reveals that statistical methods like subspace principal component analysis (Sub-PCA) dominate univariate settings, whereas deep learning architectures demonstrate superior modeling capacity in multivariate scenarios (Time Series Benchmark for Anomaly Detection (TSB-AD)-M). As shown in Figure 3.2, convolutional neural networks (convolutional neural network (CNN)) and generative models like stochastic recurrent neural network model OmniAnomaly (Omni-Anomaly) consistently outperform statistical baselines in capturing non-linear dependencies across multiple sensor channels.

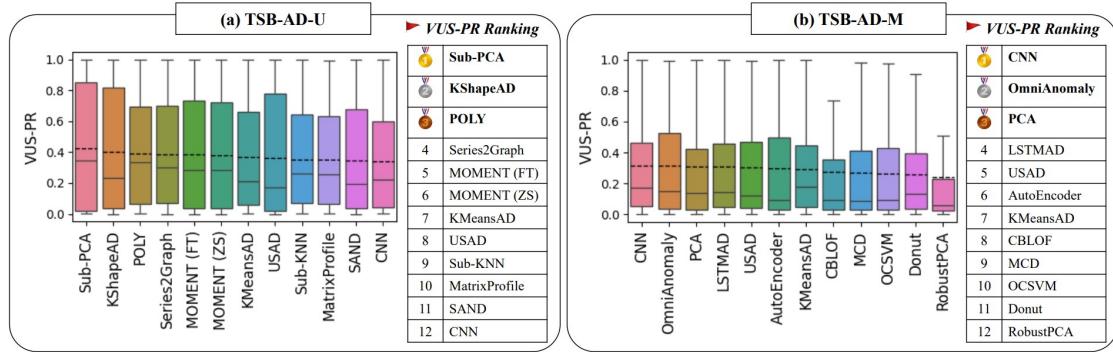


Figure 3.2: Accuracy evaluation of the top 12 methods on (a) univariate (TSB-AD-U) and (b) multivariate (TSB-AD-M) datasets based on the VUS-PR metric. Adapted from Liu and Paparrizos [LP24].

3.1.3. Implications for Multivariate Context Point Anomalies

While the **TSB-AD** benchmark provides a critical foundation for metric selection, its direct application to building energy telemetry is limited by several domain-specific gaps. The benchmark established that **machine learning (ML)** architectures like **CNN** excel in multivariate dependency modeling, while foundation models demonstrate superior efficacy in point anomaly identification. However, the **TSB-AD-M** partition contains a limited representation of multivariate point anomalies; the majority of its instances consist of sequence-based deviations or global outliers rather than contextual point anomalies.

Furthermore, Liu and Paparrizos [LP24] primarily evaluated foundation models in univariate contexts, leaving their performance in multivariate environments unexplored. This is partly due to the historical limitation of these models to univariate inputs, a constraint only recently addressed by the emergence of architectures such as *Chronos-2*, which extends universal forecasting to multivariate data [Ans+25].

For building energy systems, the benchmark lacks specific energy-sector data and does not account for the longitudinal nature of building operations. In real-world scenarios, researchers often have access to multiple years of historical data, which allows for the establishment of robust baselines. Unlike the static snapshots in many benchmarks, building data is subject to slow behavioral drifts (e.g., equipment aging). This necessitates a benchmark setup where models can continuously learn from historical patterns before being evaluated on anomalies. Consequently, while this thesis utilizes **VUS-PR** and the architectural hints provided by Liu and Paparrizos [LP24], the experimental design is adapted to leverage long-term historical training and the specific requirements of multivariate contextual detection.

3.2. Comparative Analysis of Deep Learning and Foundation Models in Energy Systems

The landscape of **TSAD** within energy data has shifted from classical statistical heuristics toward complex **deep learning (DL)** architectures and, more recently, **foundation model (FM)**. While Morshedi and Matinkhah [MM25] provide a comprehensive overview of the efficacy of **CNN**, **long short-term memory network (LSTM)**, and **generative adversarial network (GAN)** in **Internet of Things (IoT)** environments, the specific requirements of building energy telemetry necessitate a more nuanced evaluation of model topology and predictive mechanisms.

3.2.1. Deep Generative Models and the Advantage of Reconstruction

A critical distinction in building energy research involves the choice between deterministic and generative modeling. Azzalini et al. [Azz+25] conducted an empirical evaluation of deep autoencoders, demonstrating that **LSTM**-based architectures generally outperform convolutional variants due to their ability to capture long-range temporal dependencies in sequential meter data.

A significant finding in this context is the superiority of **reconstruction probability (RP)** over simple **reconstruction error (RE)**. In **Variational Autoencoder (VAE)** frameworks, **RP** accounts for the variability of the reconstruction by utilizing the learned variance, which makes the model more robust against the inherent noise of building systems. This aligns with the architecture of **OmniAnomaly**, a deep generative model that leverages stochastic recurrent neural networks to capture the normal patterns of multivariate data through robust latent representations [Su+19].

3.2.2. The Emergence of Time-Series Foundation Models

The applicability of generalized models to the energy domain remains a primary research gap. Hela, Handigol, and Arjunan [HHA25] investigated the performance of **TSFM** such as *TimeGPT* and *MOMENT*, finding that while **FM** demonstrate strong zero-shot capabilities for point anomalies, they often struggle with the domain-specific complexities of energy systems without proper adaptation.

However, the recent emergence of architectures capable of handling multivariate dependencies, such as *Chronos-2*, suggests that the limitations observed in earlier benchmarks may be overcome [Ans+25]. These models utilize a “universal forecast-

ing” approach that treats multivariate signals as tokens, allowing the model to learn cross-channel correlations that are essential for identifying context-based deviations in building energy consumption.

3.2.3. Synthesis: Prediction-Based Stochastic Modeling

A synthesis of current literature reveals a converging trend toward prediction-based methodologies as the primary solution for multivariate context point anomalies in building energy data (see Section ??). While reported results remain susceptible to the data-level deficiencies and biased metrics identified in Section 3.1.1, the scientific consensus points toward a transition from deterministic modeling to stochastically adjusted prediction. It was shown through the findings of Azzalini et al. [Azz+25], Su et al. [Su+19], and Ansari et al. [Ans+25] that robust results are achieved when a model outputs a probability distribution rather than a single point value.

This development is consistent with the established physical nature of building telemetry. As described in Section 2.1.5, energy signals are inherently stochastic, characterized by non-normal mixture distributions and multimodal profiles that originate from the interaction of discrete system states and irregular human behavior. Because traditional mean-based detection is ineffective for such data, the prediction model must inherently mimic these stochastic properties to establish a reliable normative baseline.

While advanced **FM** and generative models define the current frontier of this stochastic approach, standard **ML** architectures—specifically **CNN** variants—maintain a vital role in the field due to their verified capacity to model non-linear spatial dependencies across multiple sensor channels. Conversely, while simple statistical methods remain valuable for validation, they lack the complexity required to resolve context-dependent deviations within the multi-year historical datasets utilized in this research. This project, therefore, prioritizes architectures that utilize stochastic prediction logic. Whether implemented via the multivariate forecasting of *Chronos-2* or the deep generative reconstruction of **OmniAnomaly**, the integration of probability distributions remains the verified standard for resolving the complexities of building energy telemetry.

3.3. TODO: Mixture Density Networks for Stochastic Modeling

This section will introduce Mixture Density Networks (MDN) as a mechanism for producing probabilistic outputs in anomaly detection. It will summarize the original MDN formulation, discuss how a **CNN** feature extractor can be combined with an MDN out-

put head to model multimodal distributions in building energy data, and position this approach relative to existing generative models such as [VAE](#) and [OmniAnomaly](#).

3.4. TODO: Root Cause Analysis for Anomaly Detection

This section will review existing approaches to anomaly diagnosis and root cause analysis in time series systems. It will cover methods for attributing detected anomalies to specific sensors or subsystems, discuss graph- or topology-based RCA in building management, and examine how anomaly scores can be translated into operational and financial impact for facility managers.

3.5. Identification of Research Gaps

The current state of the art, summarized by the findings of Liu and Paparrizos [[LP24](#)], reveals that existing benchmarks and models are insufficient for the specific requirements of building energy telemetry. The subsequent sections identify the primary gaps that inform the methodology developed in this research.

3.5.1. Representation Gap in Multivariate Context Point Anomalies

A primary deficiency in existing benchmark datasets is the lack of specific representation for Multivariate Context Point Anomalies (MCPA). Most established benchmarks focus predominantly on sequence anomalies or global outliers, leaving the interaction of multivariate context point deviations under-explored. Furthermore, existing archives often lack authentic energy consumption data, utilizing synthetic or unrelated [IoT](#) datasets that do not exhibit the multimodal and stochastic characteristics inherent to building telemetry.

3.5.2. Temporal Baseline and Generalization Gap

The relationship between training data length and model performance in non-stationary building environments is currently not well-documented. Because building behavior evolves over time due to equipment degradation and seasonal shifts, it remains unclear how models trained on short-term baseline periods, such as two weeks, compare to

those utilizing long-term historical data of one year or more. Additionally, there is a lack of empirical evidence regarding the seasonal translation capability of models. Specifically, it has not been sufficiently proven whether a model trained exclusively on winter heating cycles can accurately identify anomalies during summer cooling cycles without generating excessive false positives.

3.5.3. Architectural Gap in Stochastic Mixture Modeling

While **TSFM** have shown quantifiable results in univariate forecasting, their application to multivariate energy anomaly detection represents a very recent development. The emergence of *Chronos-2* provides a technical opportunity to evaluate universal forecasting on multivariate building data. However, a significant gap remains in combining the spatial feature extraction capacity of **CNN** architectures with a stochastic output head. By integrating **CNN** variants with Mixture Density Networks (MDN), a model could output a probability distribution instead of a deterministic point value, which would align more closely with the multimodal nature of building energy data.

3.5.4. Diagnostic and Economic Functional Gap

The majority of current research in **TSAD** terminates at the detection phase, providing binary or continuous anomaly scores without further diagnostic depth. In building management, an alert without context is of limited operational utility. A critical gap exists in the automated identification of the root cause, which refers to the specific sensor or subsystem responsible for the deviation. Furthermore, there is a lack of integrated frameworks that translate detected energy waste into financial impact, defined as the monetary loss incurred by the anomaly over a specific time interval.

3.6. Synthesis of Research Objectives

The objective of this research is to fill these identified gaps by developing a custom benchmark tailored to building energy telemetry. By utilizing the **VUS-PR** metric as the evaluative standard, the methodology focuses on the detection of multivariate context point anomalies across varying training baseline lengths. The proposed approach integrates the spatial modeling success of **CNN** architectures with stochastic mixture density outputs and evaluates the multivariate capabilities of *Chronos-2*. Ultimately, the

framework aims to extend the detection pipeline into automated root cause analysis and financial impact quantification to provide actionable insights for building operations.

4

Methodology

4.1. System Context: The Eliona IoT Platform

The anomaly detection system is integrated into the Eliona IoT Platform, which serves as the operational environment for data ingestion, storage, and visualization[[Eli25b](#); [Eli25f](#)]. The platform is designed to be deployment-agnostic, operating primarily as a high-scale, Azure-based Cloud environment while preserving on-premise capability for local installations. This flexibility allows the same anomaly detection logic to be applied consistently across multiple tenants and deployment models.

4.1.1. Modular System Architecture

The platform is organized into three functional layers to ensure data isolation, scalability, and high throughput. Computational logic and specialized microservices reside within the backend, while the frontend provides a comprehensive interface for visualization and user interaction.

Device Layer This layer connects physical assets via standard protocols such as MQTT, HTTP, BACnet, and Modbus. Each device is uniquely authenticated using credentials or tokens to ensure secure and traceable data ingestion.

Server Layer (Backend) This layer acts as the centralized processing hub. It manages asset registration, hosts the Rule Engine for automated data processing, and coordinates specialized microservices for distinct use cases[[Eli25d](#); [Eli25e](#)]. Time-series data is stored in a single PostgreSQL instance extended with TimescaleDB,

using conventional relational tables for metadata and Hypertables for high-frequency telemetry.

Application and Frontend Layer This layer serves as the primary interface for end-users. It provides real-time dashboards, maps, reports, and analytics for monitoring energy health and interacting with the results produced by backend calculations.

4.1.2. Asset Modeling and Hierarchical Ontology

A central feature of the platform is its asset model and ontology, which provide a structured representation of entities and their relationships in building data [Eli25c; Eli25a]. Assets are created from reusable templates and organized into multiple hierarchies to reflect both physical layout and functional dependencies.

Assets and Templates Assets represent any entity in the system, including sensors, rooms, equipment, or entire buildings. Each asset is instantiated from an Asset Template that predefines attributes such as temperature, occupancy, or power demand, enabling consistent metadata across sites and tenants[Eli25b; Eli25a].

Dual Hierarchies Assets are structured into two complementary tree structures. The Local Tree captures physical location (e.g., Site → Building → Floor), while the Functional Tree represents technical relationships (e.g., Heating System → Pump → Flow Sensor)[Eli25f]. This dual representation enables both spatial and functional queries over the same telemetry.

Tagging Metadata tags are assigned to assets to group and query telemetry points across different buildings and tenants. Tags provide an additional semantic layer on top of the hierarchies, which is utilized by the anomaly detection system to retrieve relevant multivariate signals for model training and scoring.

4.2. System Requirements Specification

The following specifications define the mandatory capabilities of the integrated system and the detection methodology.

4.2.1. Functional Requirements

The functional requirements focus on the operational utility and the diagnostic depth of the system within the hierarchical building environment.

Multi-Tenancy and Hierarchy Management The system must facilitate the simultaneous processing of data for multiple tenants while ensuring strict data isolation between organizations. It must support a structural hierarchy where each tenant governs multiple sites, and each site contains diverse building complexes.

License-Based Activation The implementation must provide granular administrative control to activate or deactivate anomaly detection services for specific tenants based on their current license status.

Multivariate Contextual Modeling To identify context-dependent deviations, the architecture must integrate parallel sensor channels, specifically incorporating site-localized weather data to capture environmental dependencies.

Financial Impact Quantification The pipeline must automate the calculation of monetary costs associated with energy waste by quantifying the deviation between the observed consumption and the predicted normative mean.

Automated Root Cause Analysis (RCA) Detected anomalies must be traceable to their specific sensor or subsystem origin by leveraging the platform's functional tree and building ontology.

Cross-Layer Observability The system must detect faults across different operational planes, specifically identifying issues in the Supply Layer (e.g., generation failure) and the Control Layer (e.g., setpoint malfunctions), regardless of whether these issues were present during the initial baseline period.

Non-Stationary Adaptation Detection models must adapt to evolving building characteristics, such as equipment degradation or slow behavioral shifts, to maintain long-term accuracy.

Anomaly Persistence Management Mechanisms must be implemented to prevent recurring or persistent anomalies from being incorrectly integrated into the "normal" baseline, ensuring that continuous faults remain flagged as deviations.

4.2.2. Operational and Data Integrity Requirements

These requirements define the technical resilience and performance constraints necessary for high-throughput production environments.

Data Quality Resilience The ingestion and preparation layer must identify and bypass data integrity issues, such as transmission gaps and sensor recovery spikes, to prevent the generation of false-positive alerts.

Exclusion of External Drivers The system must distinguish between internal technical faults and extraordinary external events, such as extreme outdoor temperature fluctuations, to avoid incorrectly marking demand-side spikes as system anomalies.

Manual Baseline Configuration The interface must allow users to manually define "healthy" baseline periods, ensuring the model establishes a normative profile based on optimal operational states.

Scalability and Parallelism The architecture must be horizontally scalable, enabling the parallel processing of data streams across multiple buildings and sites to meet high-throughput requirements.

4.3. Proposed High-Level Architecture

The proposed system utilizes a modular, microservice-oriented design to fulfill the requirements of multi-tenancy and scalability. It is realized through a closed-loop data pipeline that bridges the existing Eliona infrastructure with specialized detection logic.

4.3.1. Component Interaction and Data Flow

The architecture consists of three primary technical domains: the Eliona Core, the Anomaly Detection Microservice, and the Analytics Endpoint.

Eliona Core (Backend and Database) This domain serves as the source of truth, providing raw telemetry via the central PostgreSQL/TimescaleDB instance.

Microservice Docker Acting as the orchestration layer, this autonomous container manages the ingestion of multi-tenant data. It performs data preparation, which

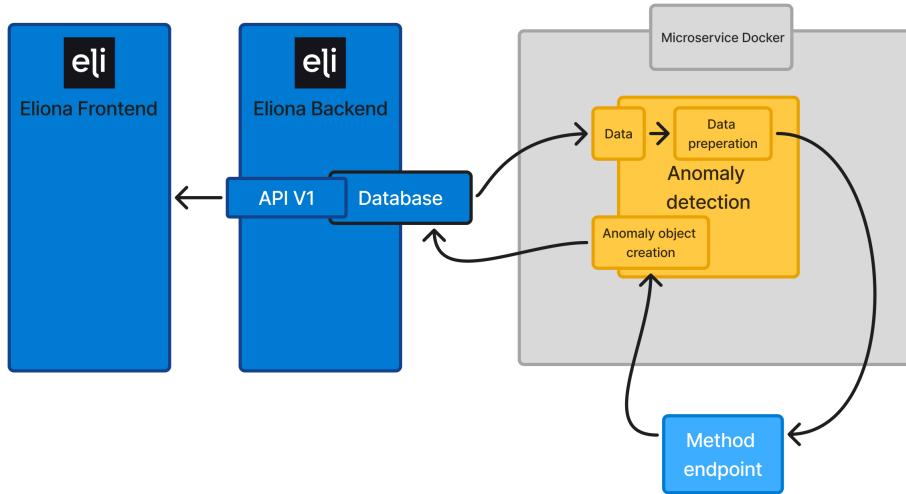


Figure 4.1: High-level architecture of the integrated anomaly detection system, illustrating the interaction between the Eliona Core, the anomaly detection microservice, and the analytics endpoint.

includes cleaning signal noise and transforming time-series data into multivariate tensors suitable for the model.

Method Endpoint This dedicated component hosts the detection algorithm. It receives the prepared data tensors and returns probabilistic scores to the microservice for final processing.

Anomaly Object Creation Upon receiving a detection signal, the microservice generates a structured anomaly object. This object, containing diagnostic data and financial impact, is written back to the Eliona database and subsequently visualized in the frontend via API V1.

4.3.2. Integration into the Azure Ecosystem

For cloud-based deployments, the architecture is embedded within Azure Kubernetes Services (AKS). This ensures that the microservice can scale horizontally to handle high-throughput requirements across multiple tenants. The use of Azure Database for PostgreSQL provides a robust, distributed foundation for the platform's metadata and telemetry.

4.4. Proof of Concept (PoC): Stochastic Feasibility

The PoC was implemented to validate the hypothesis that stochastic modeling can effectively address the non-normal distributions found in building energy telemetry. This iteration focused on the core detection mechanism and the integration of diagnostic layers.

4.4.1. MDN-Based Stochastic Prediction

The initial implementation utilized a Mixture Density Network (MDN) as the core detection logic. Unlike deterministic models, the MDN was trained to output a conditional probability distribution of consumption based on contextual features such as outdoor temperature and occupancy. By minimizing Negative Log-Likelihood (NLL), the model captured the multi-modal nature of the energy signals.

4.4.2. Diagnostic Integration: SHAP and LLM

To fulfill the requirements for Root Cause Analysis (RCA) and actionable insights, the PoC incorporated two diagnostic layers:

Feature Interpretability SHAP values were utilized to identify the specific contextual drivers influencing each prediction. For instance, if the occupancy feature exhibited high impact during a night-time anomaly, it indicated a likely fault in the lighting or HVAC scheduling.

Natural Language Synthesis The results from the SHAP analysis and the financial residuals were fed into a Large Language Model (LLM). This allowed the system to generate human-readable explanations and remediation steps, transforming abstract data into operational guidance.

4.5. Critique of Sequential Forecasting for Anomaly Detection

A dominant paradigm in time-series anomaly detection is the use of sequential forecasting models. In this approach, a model (e.g., RNN, LSTM, or Transformer) is trained to predict the next value x_t based on a sliding window of historical values $(x_{t-w}, \dots, x_{t-1})$ and potentially exogenous features. An anomaly is flagged if the deviation (residual)

between the predicted value \hat{x}_t and the actual value x_t exceeds a threshold. While intuitively appealing, this autoregressive approach suffers from fundamental limitations when applied to sustained anomalies in industrial settings, particularly regarding error propagation and signal adaptation. To demonstrate these failure modes, a controlled synthetic experiment was conducted.

4.5.1. Synthetic Experimental Setup

A synthetic dataset was generated to simulate a predictable building energy profile: consumption is set to 10 units between 08:00 and 18:00 on weekdays, and 0 units otherwise. To evaluate detection capabilities, two distinct, sustained anomalies were injected:

1. A “night-shift” anomaly with sustained consumption of 10 units during nighttime hours.
2. A “weekend-work” anomaly with sustained consumption of 5 units over a weekend.

Three distinct forecasting models, plus an additional inference-time variant of the 24-hour model, were tested against this data to highlight different behavioral modes. The anomaly score is calculated as the absolute difference between actual and predicted values.

4.5.2. Failure Mode 1: Error Propagation and Instability

The first fundamental issue arises when a sequential model encounters substantial, previously unseen anomalous data. Because the model relies on past observations to generate future predictions, once an anomaly occurs, it enters the model’s input window for the subsequent w steps.

Figure 4.2 illustrates this phenomenon using a model trained with a 24-hour historical window plus time-based features (time of day, `is_weekend`).

When the sustained nighttime anomaly hits, it represents data completely outside the model’s training distribution. The model fails to predict the onset (generating a high anomaly score initially). However, as these anomalous 10-unit values fill the 24-hour input window, the model’s internal state becomes corrupted. It begins making erratic predictions, sometimes overestimating, sometimes underestimating, resulting in a noisy anomaly score signal. Crucially, this instability persists even after the actual anomaly has finished, as the “poisoned” window takes 24 hours to clear.

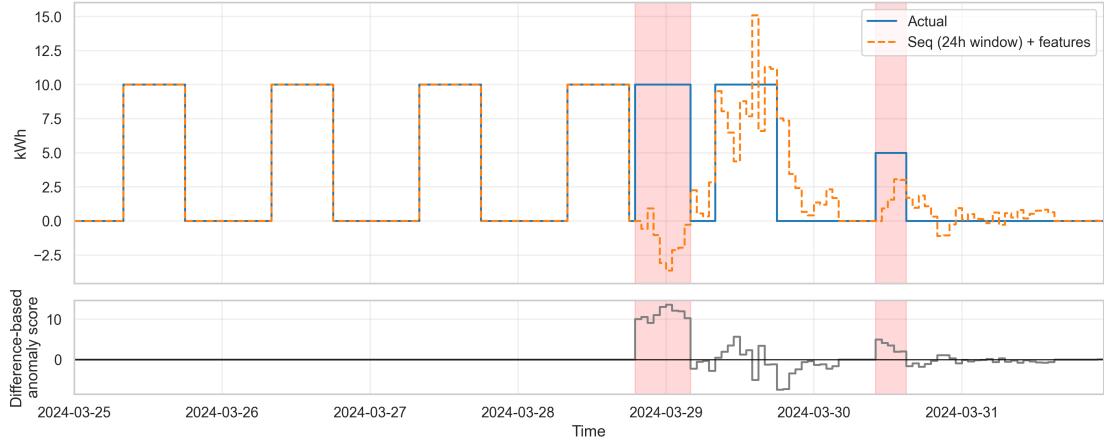


Figure 4.2: Prediction behavior of a model using a 24 h historical window plus time features. The top panel shows actual vs. predicted values; the bottom panel shows the difference-based anomaly score. Note the erratic predictions even after the anomaly ends as the unseen data propagates through the sliding window.

4.5.3. Failure Mode 2: Rapid Adaptation and the PA-F1 Illusion

The second failure mode is conversely related to models relying heavily on short-term autocorrelation. In many time series, the best predictor of x_t is simply x_{t-1} . If a model learns this dependency strongly, it will rapidly “adapt” to a sustained anomaly.

Figure 4.3 shows a model trained only on the past five historical values, without contextual features.

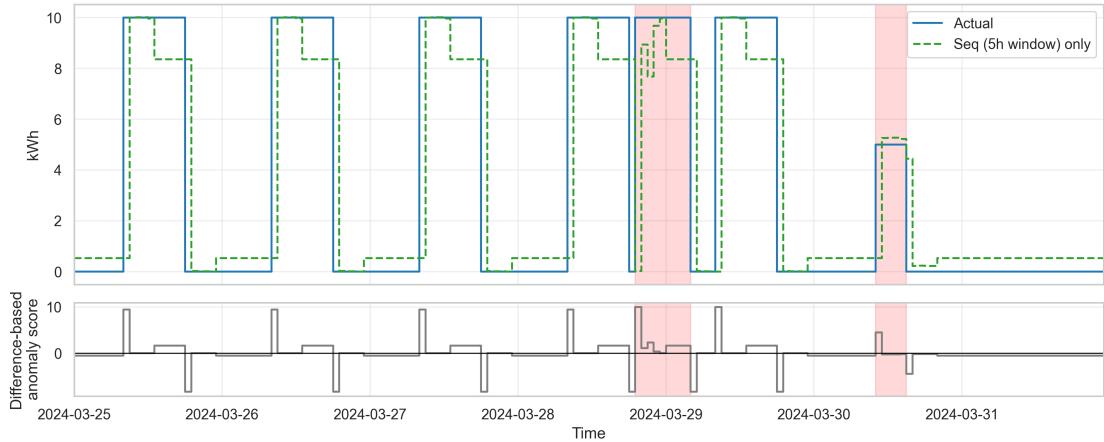


Figure 4.3: Prediction behavior of a model using only a short (5-step) historical window. The model correctly identifies the onset of anomalies but rapidly adapts to the new level, causing the anomaly score to drop back to near zero while the anomaly is still ongoing.

The model successfully flags the onset of both anomalies due to the sudden jump. However, within five time steps, the input window is filled with the anomalous values. The model quickly learns the new “normal” (e.g., that consumption is currently 10 at

night) and predicts accordingly. The residual drops to near zero, and the anomaly is effectively missed for the majority of its duration.

Implications for Evaluation Metrics and Financial Impact

This behavior explains the heavy reliance in academic literature on Point Adjustment F1 (PA-F1) scores. In PA-F1, if a model detects a single point within a contiguous anomaly segment, the entire segment is counted as correctly detected. While this inflates benchmark scores, it masks the model's inability to track sustained deviations.

For industrial applications requiring financial impact quantification, this failure mode is catastrophic. Calculating financial loss requires integrating the deviation over the entire duration of the event. A model that only flags the first 15 minutes of a 4-hour energy spike is useless for quantifying the total wasted energy.

4.5.4. Mitigation Strategies

There are two primary architectural strategies to resolve these sequential dependence issues.

Strategy A: Contextual Feature-Only Modeling

The most direct solution is to remove the autoregressive dependency entirely. By training a model to predict consumption based solely on contextual features (time, weather, occupancy) and ignoring past consumption values, error propagation is impossible.

Figure 4.4 demonstrates this approach. The prediction remains stable regardless of the actual input, providing a clean, continuous anomaly score throughout the duration of both events. While highly effective for context anomalies, this approach sacrifices the ability to model complex temporal dynamics and cannot leverage powerful sequential foundation models.

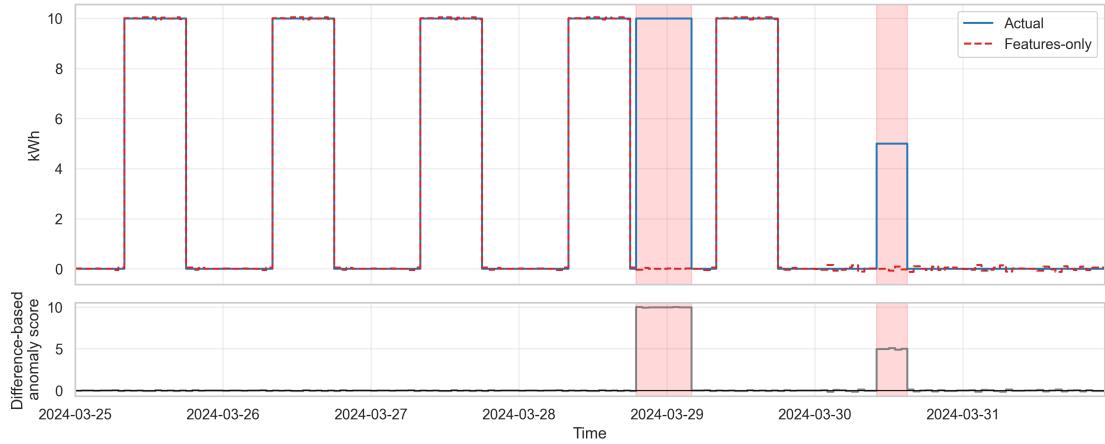


Figure 4.4: Behavior of a features-only model (no historical consumption input). The prediction relies solely on context (time/weekend), resulting in a stable baseline and accurate detection of sustained anomalies without adaptation.

Strategy B: Inference-Time Input Imputation

To retain the benefits of sequential modeling while mitigating error propagation, an inference-time correction mechanism can be introduced. If the anomaly score at step t exceeds a defined threshold, the actual value x_t is considered contaminated. Instead of feeding x_t into the sliding window for step $t+1$, the model's own prediction \hat{x}_t is imputed as a “corrected” value. In an online or periodically retrained setting, this also prevents the model from adapting its baseline to these anomalous segments, so similar future events are not reinterpreted as normal behaviour despite the non-stationarity of the raw building signal.

Figure 4.5 applies this logic to the unstable 24-hour window model from Figure 4.2. By replacing anomalous inputs with predictions, the sliding window remains clean, preventing the model from adapting to the anomaly or becoming unstable. This allows for accurate tracking of sustained anomalies while still using sequential architectures.

4.6. Statistical Limitations of Point and Gaussian Predictions

To isolate the effect of distributional assumptions on anomaly detection, a synthetic “Variable Shift” dataset was created. Each hourly sample toggles between a low-power regime (0–1 kWh) and a high-power regime (9–11 kWh) with a stochastic morning/evening schedule (approximately 60/40 split). A stuck-at fault of 5 kWh was injected during a regular weekday to emulate a latent control failure. Figure 4.6 shows that all three model families—deterministic dense regression, single-Gaussian prediction, and Mixture Density Networks (MDN)—deliver visually similar means, yet their anomaly

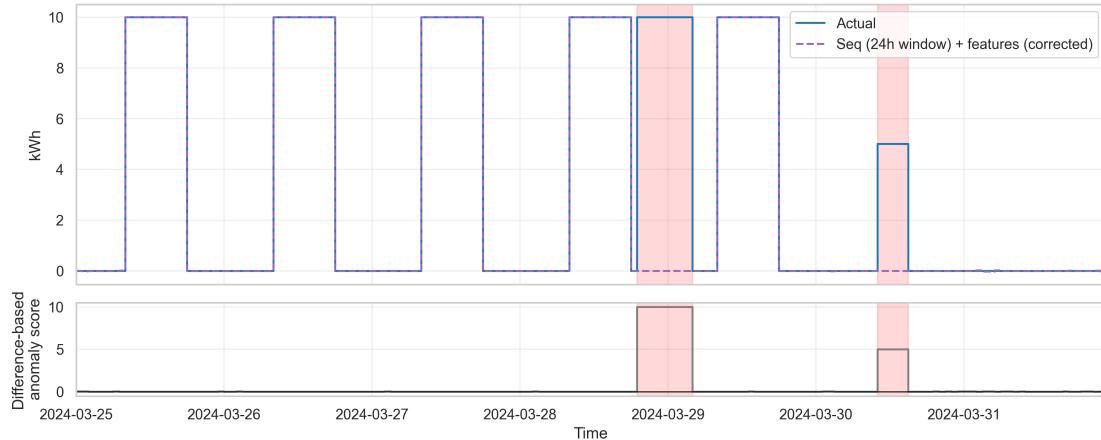


Figure 4.5: The same 24-hour window model from Figure 4.2, but applied with inference-time imputation. When an anomaly is detected, the predicted value replaces the actual value in the sliding window for future steps. This prevents error propagation and maintains a high anomaly score throughout the event.

scoring behavior diverges drastically.

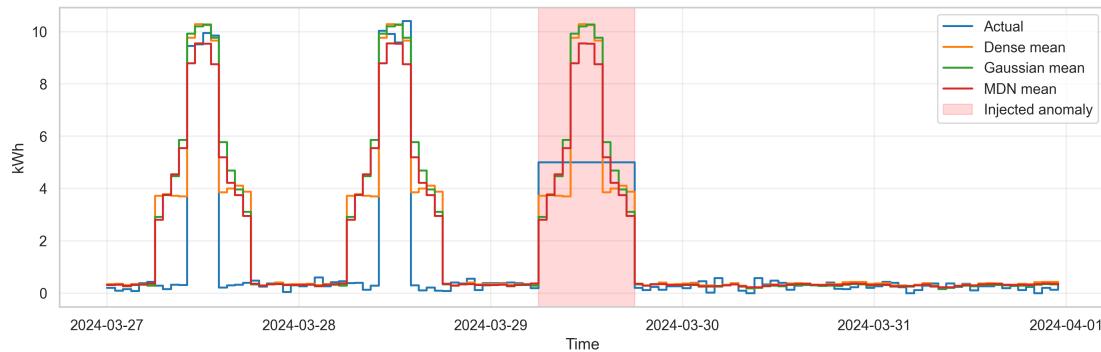


Figure 4.6: Predicted means over the Variable Shift horizon. Dense, Gaussian, and MDN models track the two regimes, masking the scoring deficiencies discussed in Sections 4.6.1–4.6.3.

4.6.1. The Failure of Mean Squared Error Minimization

Dense regressors trained with Mean Squared Error (MSE) converge toward the global average of both regimes. In bimodal settings this leads to systematic bias: the model predicts approximately 5 kWh regardless of whether the system is in its “Off” (low) or “On” (high) state. Consequently, perfectly normal behavior is scored as highly anomalous, whereas the injected stuck-at-5 event appears deceptively healthy because it matches the biased mean. The residual trace in Figure 4.7 exposes this contradiction: the absolute error balloons whenever the device operates normally, yet it contracts when the genuine anomaly occurs.

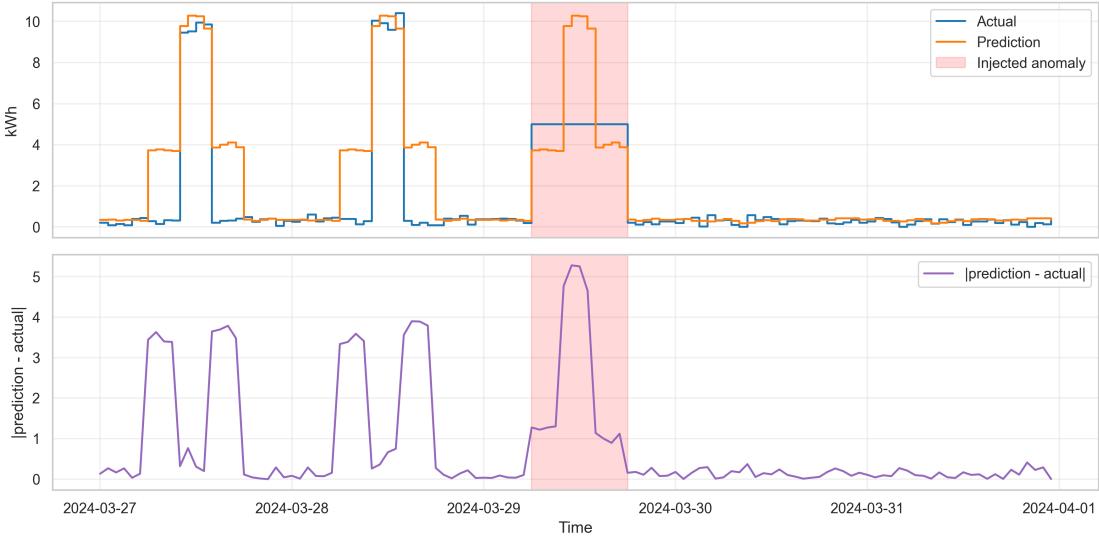


Figure 4.7: Dense regressor residuals over the Variable Shift dataset. The mid-range prediction inflates anomaly scores for legitimate operating states, while the stuck-at-5 fault yields a small residual.

4.6.2. The Gaussian Distribution Paradox

A single-component Gaussian attempts to reconcile bimodality by inflating its variance. The resulting Probability Density Function (PDF) concentrates probability mass near the center—a region never visited by real data. The normalized log-likelihood trace (Figure 4.8) confirms that the stuck-at-5 anomaly sits inside the “most likely” area of the Gaussian, generating a low penalty. Meanwhile, legitimate regime values land in lower-density shoulders and spuriously raise the score. The heatmap in Figure 4.9 makes the distortion visible: the green, high-probability band spans the median instead of the true modes.

4.6.3. Solution: Mixture Density Networks

Mixture Density Networks address both issues by learning multiple kernels simultaneously. Each component can specialize in a particular operating mode, while the regions between components retain near-zero probability. Figure 4.11 shows how the MDN assigns green (high probability) bands only where data is observed, keeping the mid-range red. When log-likelihood is used as the anomaly score, the stuck-at-5 fault immediately falls into the valley between components, producing a sharp increase in $|\log p(x)|$ (Figure 4.10). This probabilistic separation allows the MDN to quantify financial impact reliably: integrating the residual energy over time now reflects the true magnitude of the fault rather than artifacts of model bias.

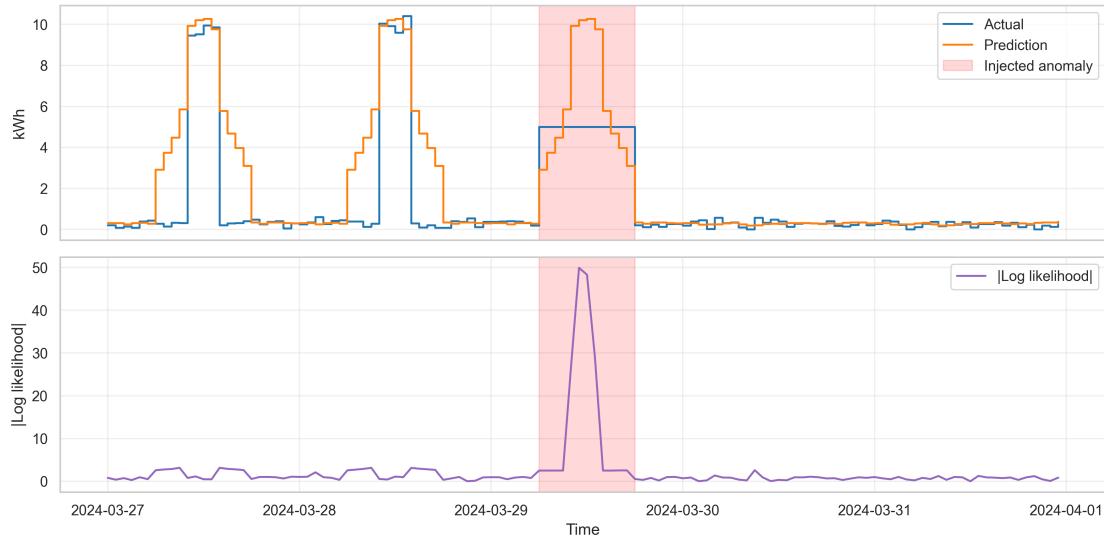


Figure 4.8: Absolute log-likelihood trace for the single-Gaussian predictor. The stuck-at-5 anomaly aligns with the high-likelihood center, suppressing the score.

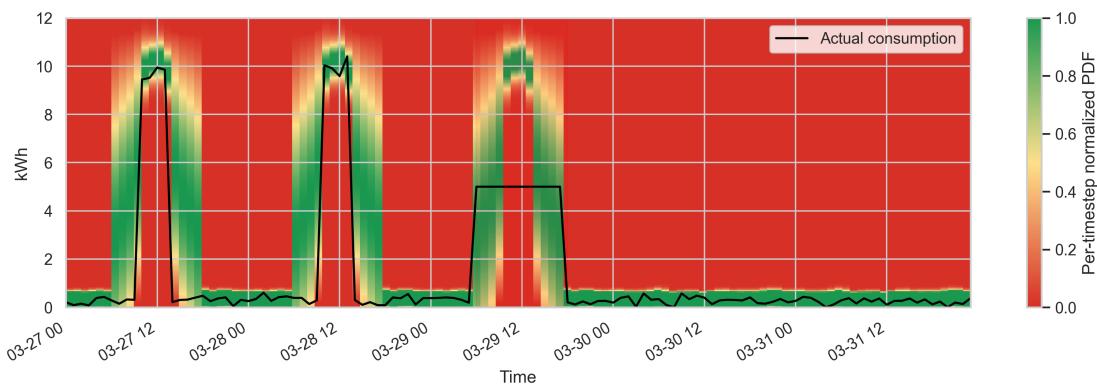


Figure 4.9: Per-timestep normalized PDF for the Gaussian model. High probability mass accumulates between the actual clusters, illustrating the variance-stretching paradox.

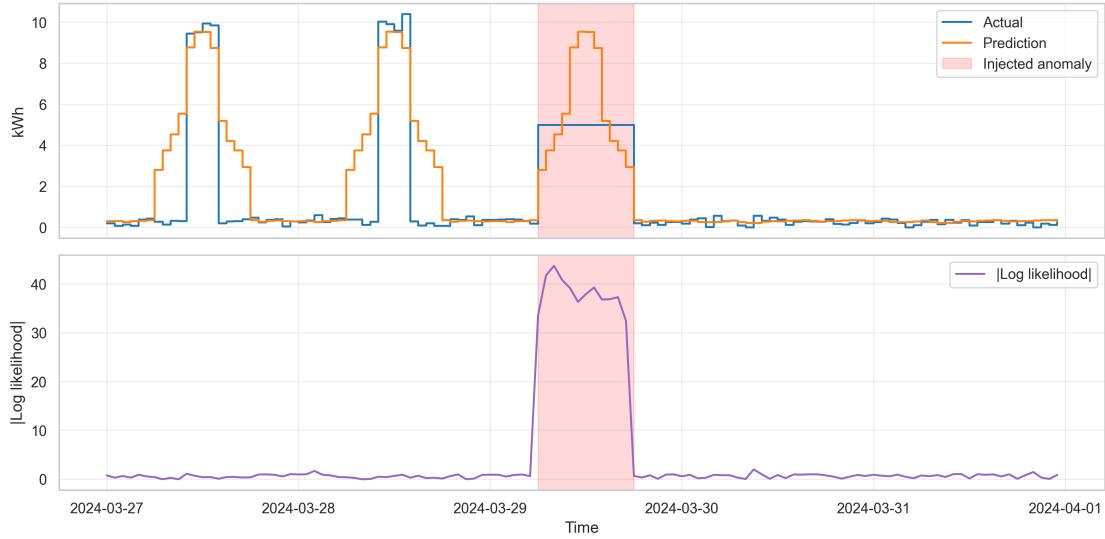


Figure 4.10: MDN absolute log-likelihood trace. The stuck-at-5 anomaly triggers a sustained spike because the value resides in a low-probability region between mixture components.

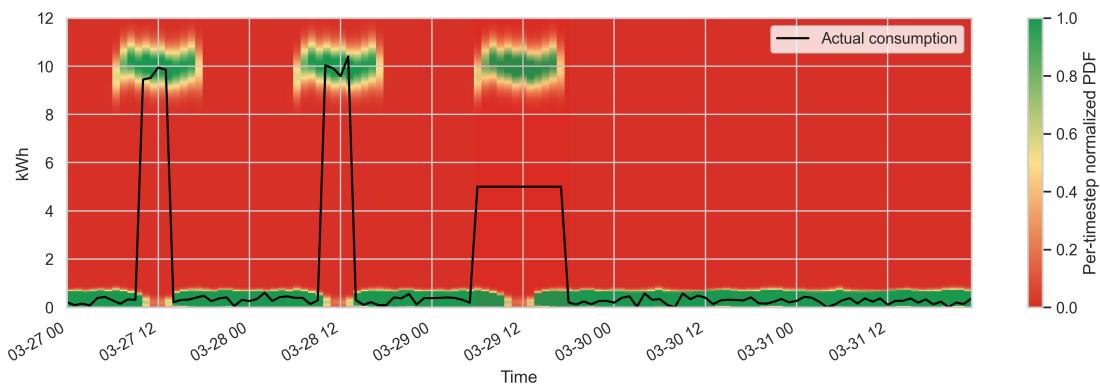


Figure 4.11: MDN normalized PDF heatmap. Two distinct high-probability ridges align with the real operating modes, while the middle band remains improbable.

4.7. Benchmark Data Generation and Composition

To compare the proposed methodology against established baselines, a comprehensive benchmark dataset was synthesized with the Building Optimization Performance Test Framework (BOPTEST)[**BOPTEST**]. BOPTEST provides high-fidelity building simulations that include weather, occupancy, and HVAC subsystems, offering a controlled sandbox for evaluating diagnostic strategies without legacy noise.

4.7.1. Simulation Environment and Baseline Construction

The “Multizone Office Complex Air” test case was selected as it emulates a large office with coupled thermal zones, AHUs, and realistic schedules.

Data fidelity The simulator injects geographically consistent weather files, dynamic occupancy, and holiday calendars, ensuring the external drivers mirror commercial buildings.

Integrity Because all sensors are generated virtually, the baseline year is free of telemetry dropouts, frozen devices, or undocumented setpoint overrides, giving a clean reference profile for “healthy” operation.

Temporal scope A continuous year of data (15-minute granularity) was exported to capture seasonal shifts and enable out-of-season evaluation.

4.7.2. Feature Selection and Control Layer Logic

The raw export was decomposed into contextual features and consumption targets while deliberately excluding control-loop variables that could leak fault signatures.

Leakage mitigation Control-layer signals (e.g., valve positions, supply setpoints) were omitted. If a stuck valve drives excess consumption and its command signal is provided to the model, the anomaly becomes “explainable” by the feature set and disappears in the residual.

Included features Exogenous drivers such as outdoor temperature, global horizontal irradiance, wind speed, occupancy counts, calendar flags (weekday/weekend, holidays), and cyclical encodings (hour-of-day, day-of-year).

Target variables Seventeen metered electricity channels including the aggregate main meter, chiller and boiler production meters, AHU fan feeds, pump circuits, and

lighting zones. This enables both whole-building financial attribution and subsystem diagnostics.

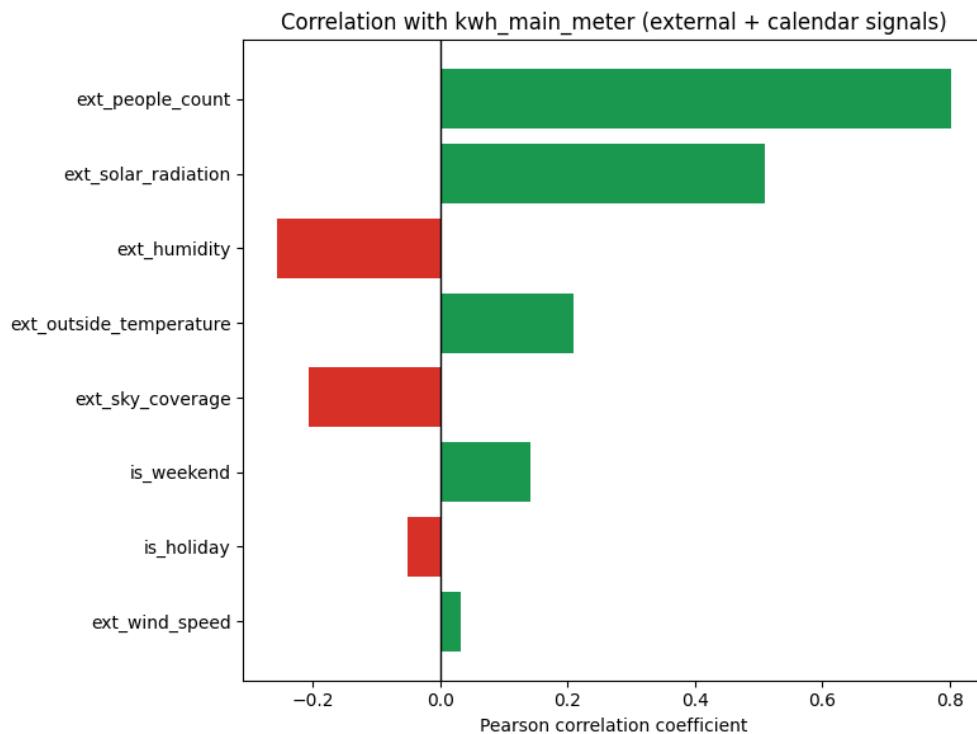


Figure 4.12: Empirical Pearson correlation of contextual and external drivers with the main electricity meter over the synthetic 2007 benchmark year.

4.8. Experimental Data Segmentation and Anomaly Injection

To probe generalization and sample efficiency, the baseline year was sliced into multiple training/testing regimes and augmented with a bounded taxonomy of synthetic faults. Anomalous points never exceed 5% of any derived dataset to preserve realism on the aggregated main meter.

4.8.1. Segmentation Strategy

- **Long-term:** Two non-overlapping six-month windows to analyze seasonal translation (e.g., winter learning, summer detection).
- **Medium-term:** Four quarterly (three-month) windows to test model freshness

requirements.

- **Short-term:** Four two-week slices representing spring, summer, fall, and winter for low-data scenarios.

4.8.2. Anomaly Taxonomy and Labeling

Custom perturbations were injected per segment, spanning operational, control, and contextual deviations. Table 4.1 summarizes the categories applied to the main meter and selected subsystems.

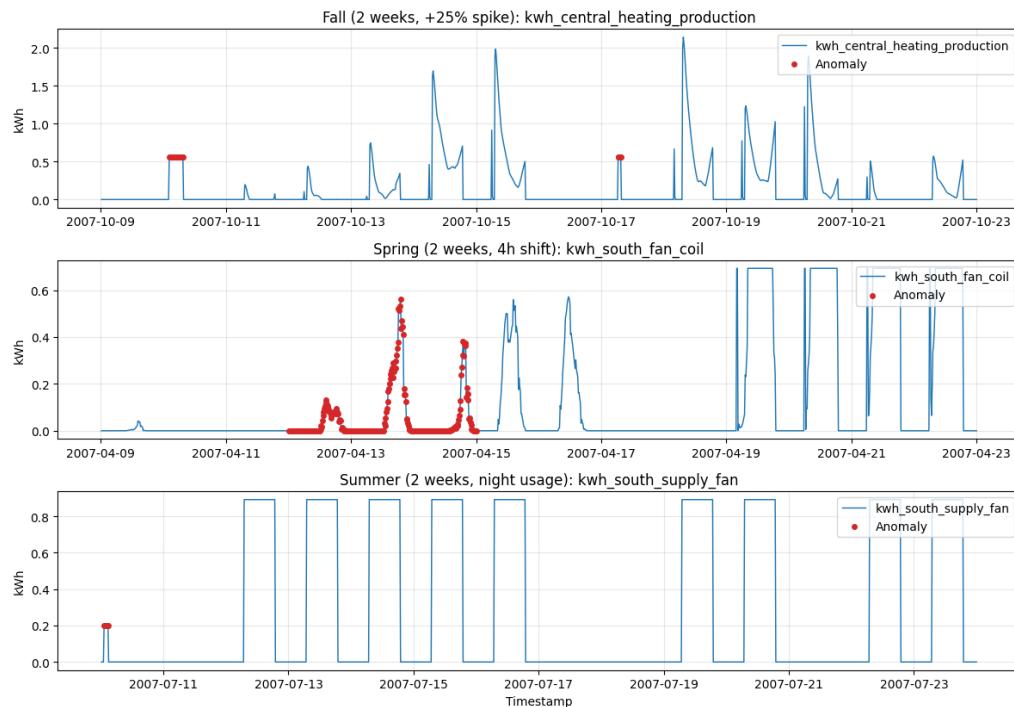


Figure 4.13: Representative sub-meter excerpts from the curated benchmark slices (fall spike, spring pattern shift, summer off-hours). Red markers indicate anomaly windows where the targeted device deviates from its baseline regime.

Each injected anomaly is labeled with its category, onset, and magnitude, enabling fine-grained evaluation of detection latency, duration coverage, and financial attribution across the benchmark suites described above.

Table 4.1: Synthetic anomaly taxonomy used in the BOPTEST benchmark datasets.

extbfCategory	Description
Device failures	Full outage (0 kWh) and degraded operation (50% load) for selected meters.
Stuck states	Forced high-load and low-load plateaus irrespective of schedule.
Drift profiles	Exponential and linear drifts (10%–60%) applied over multi-day horizons.
Extended spikes	Sustained spikes at 10%, 25%, 50%, and 100% above baseline for varying durations.
Contextual/off-hours	Night or weekend consumption persisting at abnormal levels (high/low variants).
Pattern shifts	Phase shifts of daily load shapes by 30 minutes up to 4 hours.
Point anomalies	Single-timestep spikes from 10% to 500% of nominal demand.

5

Implementation

The previous chapter established the methodological framework and the operational requirements for the anomaly detection system. This chapter details the technical realization of these concepts, focusing on the software stack, the integration into the Azure ecosystem, and the internal orchestration logic of the Scala microservice and the Python analytics endpoint.

5.1. Integrated System Architecture and Technology Stack

The implementation utilizes a polyglot approach to leverage the specific strengths of different programming paradigms. While the orchestration and data processing are handled by a Scala microservice to ensure high-performance parallelism, the predictive modeling is realized through a Python endpoint optimized for machine learning.

5.1.1. Deployment and Cluster Integration

The system is designed as a modular Docker container that operates within the same Azure Kubernetes Services (AKS) cluster as the core Eliona microservices. This co-location ensures low-latency communication between the detection logic and the primary data storage. The architecture remains environment-agnostic; for on-premise deployments, the entire stack, including the analytics endpoint, can be executed locally as a suite of interconnected containers.

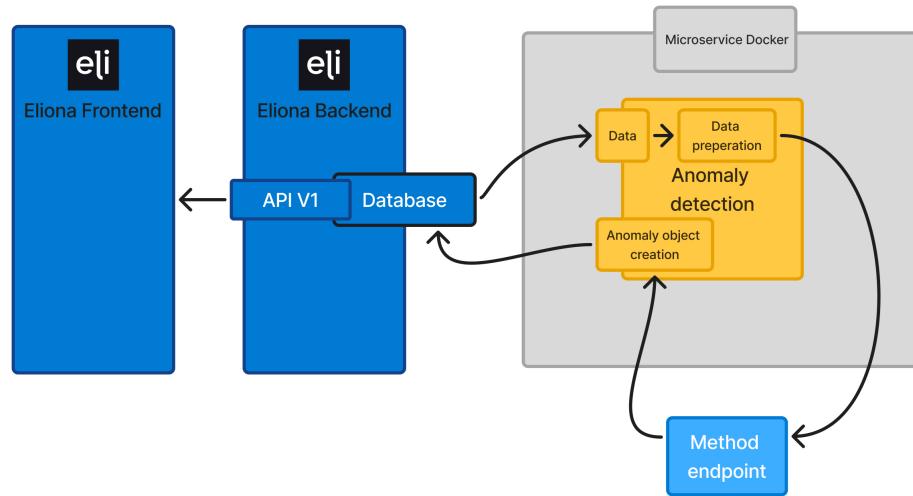


Figure 5.1: High-level deployment of the anomaly detection microservice and Python analytics endpoint within the Eliona and Azure ecosystems.

5.1.2. Data Orchestration and Persistence

The microservice establishes a closed-loop data pipeline with the Eliona backend:

Ingestion Telemetry data is retrieved directly from the centralized PostgreSQL/TimescaleDB instance.

Analytics loop After performing data preparation, the microservice transmits multivariate tensors to the method endpoint.

Synthesis and storage If the returned predictions indicate a deviation, the service creates an anomaly object. These objects are persisted in the database, where they become accessible to the Eliona frontend via API V1 for visualization and reporting.

5.2. Python Analytics Endpoint: Chronos-2 Integration

The analytical core of the system is realized as a specialized Python service. This choice is motivated by the maturity of the Python ecosystem regarding transformer-

based time-series models and the availability of managed deployment infrastructures.

5.2.1. Predictive Logic and Model Hosting

The endpoint hosts the Chronos-2 model, which is utilized in a batch prediction mode to process multiple time series simultaneously. It receives JSON payloads containing historical context and exogenous covariates, returning probabilistic forecasts and specific quantiles that are used to establish the normative operational band.

5.2.2. Managed Online Endpoints in Azure ML

For cloud-based production environments, the service is deployed as an Azure Machine Learning (AML) managed online endpoint. This infrastructure provides several quantifiable advantages:

Automated scaling The platform manages horizontal scaling and load balancing to handle variable request volumes from the Scala backend.

MLOps integration Azure ML provides a robust environment for creating training pipelines and registering versioned models, ensuring the non-stationary adaptation requirement is fulfilled.

Flexibility Despite being managed in the cloud, the service is fully containerized, allowing for a seamless transition to local Docker environments when on-premise capability is required.

5.3. Scala Microservice: Multi-Tenant Orchestration

The orchestration of the anomaly detection system is managed by an autonomous Scala microservice, which serves as the primary computational engine for data retrieval, transformation, and multi-tenant isolation. While the predictive logic is delegated to a Python-based endpoint, the Scala service functions as the “brain” of the architecture, coordinating complex asynchronous I/O operations and ensuring thread-safe execution across concurrent processing loops. The selection of Scala is motivated by the requirement for high-performance parallelism and strong static typing, which facilitates the management of the intricate building ontology and multivariate telemetry tensors. By executing on the Java Virtual Machine (JVM), the service achieves industrial-grade

scalability and resilience, allowing the system to process data from thousands of assets simultaneously without the execution bottlenecks typical of interpreted languages.

5.3.1. Multi-Tenant Lifecycle Management

The implementation ensures strict separation between different organizations through a dedicated multi-tenant management layer. Multi-tenancy is defined here as an architectural paradigm where a single instance of the software serves multiple distinct customers, known as tenants, while maintaining logical isolation of their data and configurations.

The lifecycle of these tenants is managed by the `TenantService`, which identifies licensed entities via the `DatabaseTenantRepository`. To maintain synchronization with the current platform state, the `TenantRegistry` performs a reconciliation sweep every 15 minutes. During this cycle, the registry identifies new licenses to instantiate corresponding `TenantWorkerLoop` instances, restarts workers that have encountered critical failures, and terminates processes for tenants whose licenses have expired or were removed.

Each tenant is assigned an independent `ADWorker`, ensuring that processing tasks for one organization do not interfere with the resource allocation of another. This isolation is further reinforced at the database level by a central registry that maps tenant identifiers to specific data scopes, ensuring that the `AttributeDataProcessor` only retrieves telemetry and hierarchical metadata belonging to the respective tenant. The registration of shutdown hooks ensures that all active workers, schedulers, and database connection pools are closed gracefully upon service termination, preventing data corruption or leaked resources.

5.4. Data Acquisition and Processing Pipeline

The transformation of raw building telemetry into structured inputs for the Chronos-2 model requires a multi-stage pipeline. This process ensures that the detection logic focuses on relevant energy signals while maintaining high data quality through automated cleaning and contextual enrichment.

5.4.1. Attribute Filtering and Hierarchical Selection

The system utilizes the Eliona attribute metadata to identify relevant telemetry points. To focus on electrical energy consumption, the `AttributeDataProcessor` queries all attributes assigned to the “Energy” type with *kWh* as the defined unit. To optimize computational resources and ensure operational relevance, the system applies two primary filters:

Relevance thresholding The processor analyzes the historical peak daily consumption for each meter. If the highest recorded energy consumption over a 24-hour period represents a financial value lower than a configurable threshold (standardized at \$1.00), the attribute is excluded from both the detection pipeline and the Root Cause Analysis (RCA).

Depth-limited detection While all eligible attributes are organized into a tree structure based on their physical location, anomaly detection is strictly performed on the first two layers (typically the main meter and floor meters). The remaining descendant nodes in the hierarchy are utilized exclusively for diagnostic purposes during the RCA phase.

5.4.2. Contextual Enrichment: Site-Localized Weather

The system integrates exogenous environmental data to capture the impact of external drivers on building energy demand. Each meter tree is associated with a specific site, which provides precise geographical coordinates (latitude and longitude). The `WeatherProvider` utilizes these coordinates to fetch localized weather telemetry from the Open-Meteo API. This data is persisted in the database to serve as covariates for the `ADPipeline`, ensuring that consumption spikes caused by extreme weather are not incorrectly flagged as technical faults.

5.4.3. Data Cleaning and Gap-Resilience Logic

The microservice retrieves aggregated telemetry from the database for five distinct time-frames: 15-minute, 1-hour, 4-hour, 1-day, and 1-week buckets. To ensure data quality resilience, the following preparation steps are applied to the raw signal:

Gap filling and spike redistribution Missing data buckets are filled with zero values. If a transmission gap is followed by a recovery spike—a common occurrence in

IoT gateways—the accumulated energy value is distributed equally across the duration of the preceding gap to prevent false-positive alerts.

Integrity flagging For every data point, the system generates a binary feature that is set to 1 if a gap was present and 0 if the data was received normally. This provides the model with explicit context regarding signal reliability.

Recursive imputation To prevent known anomalies from “poisoning” the sliding input window of the Chronos model, any value previously identified as anomalous is replaced with its corresponding predicted median value before being used for subsequent forecasts.

5.4.4. Reactive Data Fetching

The pipeline operates on a reactive schedule to synchronize with the database’s aggregation cycles. Every 15 minutes, a new fetch operation is triggered; however, data is only retrieved for completed aggregation buckets. For example, 15-minute buckets are refreshed every quarter-hour, whereas 1-day buckets are only updated every 24 hours once the full daily period has concluded. This ensures that the model always operates on finalized, consistent telemetry.

5.5. Stochastic Inference and Anomaly Quantification

The core of the detection logic resides in how data is transmitted to the Chronos-2 endpoint and how the resulting stochastic forecasts are interpreted to identify and quantify anomalies.

5.5.1. Feature-Driven Prediction Strategy

To fulfill the methodological requirement of reducing dependence on autoregressive lags, the system utilizes a specific inference strategy. As established in Section 4.5, standard sequential models often “adapt” to anomalies because the anomalous value becomes a feature for the next prediction. To mitigate this, the microservice does not send the actual value of the timestamp currently being evaluated for an anomaly.

Instead, the service transmits a historical buffer along with contextual features (weather, occupancy, and temporal indicators) and requests a prediction horizon that extends up

to the point of interest. By providing the model with “future” known covariates while withholding the actual consumption at the target timestamp, the model is forced to rely on the established contextual relationships rather than the most recent (potentially anomalous) observation.

5.5.2. Batch Processing and Quantile Requests

The ADPipeline batches requests for multiple attributes and timeframes into single JSON payloads to maximize the throughput of the Azure ML endpoint. For each attribute, the service requests specific quantiles from the Chronos-2 model:

- **99th percentile** ($q_{0.99}$): used as the high-sensitivity threshold for detecting extreme deviations.
- **68th percentile** ($q_{0.68}$): used as the normative baseline for financial impact calculations in high-consumption scenarios.
- **50th percentile** ($q_{0.50}$): the median, serving as the central tendency of the prediction distribution.

5.5.3. Detection Logic and Financial Quantification

An anomaly is formally triggered if the actual value x_t falls outside an extreme prediction interval defined by the 99th percentile:

$$x_t > q_{0.99,\text{upper}} \quad \text{or} \quad x_t < q_{0.01,\text{lower}}. \quad (5.1)$$

Once an anomaly is confirmed, the system calculates the financial impact. A critical challenge in multi-modal building data is that the mean often represents an improbable “middle ground” between an *On* and *Off* state. To prevent inflated impact values, the system calculates the residual distance from the 68th percentile rather than from the mean.

Negative impact (waste) If the consumption is higher than the upper bound, the wasted energy is quantified as the distance to the $q_{0.68}$ quantile. This ensures that if the normal behaviour at that time could have been a high-load state, the anomaly is measured against that high-state boundary rather than a lower global average.

Positive impact (savings) If the consumption is lower than the lower bound, the savings are measured relative to the corresponding lower quantile.

These residuals are then multiplied by the tenant-specific `energyCostPerKwh` to generate a signed monetary value, which is persisted as part of the anomaly object. To avoid cluttering the system with events that correspond to only a few cents of effect, any candidate anomaly whose absolute financial impact falls below a configurable minimum threshold is discarded and not promoted to a persisted anomaly object.

5.6. Hierarchical Root Cause Analysis (RCA)

When an anomaly is confirmed on a high-level meter, such as the main meter, the `ADWorker` initiates a targeted diagnostic sweep of the descendant hierarchy. The service identifies all sub-meters within the functional tree and retrieves their corresponding time-series trends.

5.6.1. Diagnostic Attribution

These sub-meter trends are processed through the `ADPipeline` in a specialized diagnostic mode. The system calculates the individual financial impact for each sub-component to determine their relative contribution to the primary anomaly. The resulting diagnostic data includes:

Asset-specific impact A breakdown of financial loss attributed to individual assets (for example, “Light 1: -\$3.00”, “HVAC Pump: +\$1.00”).

Asset type aggregation Contributions grouped by category to identify systemic issues, such as a specific percentage of the total impact being caused by “Lighting” or “Plug Loads”.

5.6.2. Localization and Weather Context

The diagnostic summary is enriched with localized metadata and environmental context. The system utilizes the `WeatherProvider` to retrieve the meteorological conditions at the exact time of the event, such as the outside temperature and sky coverage. This information is formatted into a standardized diagnostic string, for example: “Unusual high energy consumption on Monday at 04:30. Outside temperature 12 °C, Weather: Clear Sky.”

5.7. Temporal Collapse and Persistence

The system monitors telemetry across five distinct timeframes (from 15-minute to 1-week buckets), which frequently results in overlapping detection events. To prevent redundant alerting and maintain data integrity, the ADPipeline implements a temporal collapse logic:

Hierarchical merging Anomalies detected on shorter timeframes are absorbed into higher-timeframe events as they mature. For example, four individual 15-minute anomalies are deleted once a corresponding 1-hour anomaly is confirmed, with their diagnostic data and timestamps merged into the 1-hour object.

Impact optimization To ensure the most realistic representation of waste, the system retains the highest financial impact value between the aggregated lower-timeframe residuals and the single higher-timeframe calculation.

5.8. AI Synthesis and Recommended Actions

For anomalies exceeding a defined priority or financial threshold, the diagnostic payload is transmitted to the AnomalyExplainer. This component utilizes a Large Language Model (LLM) to synthesize the raw diagnostic data into operational insights.

5.8.1. Scenario A: Behavioral Fault (Lighting)

If the root cause analysis identifies that approximately 90% of a nighttime anomaly was caused by lighting assets, the LLM generates a targeted explanation:

Possible explanation: Technical staff or occupants likely left the lighting systems active during non-operational hours. Recommended action: Manually deactivate the identified lighting circuits and implement an automated “All-Off” logic using the platform’s rule chains.

5.8.2. Scenario B: Technical Fault or Misuse (Plug Loads)

In cases where a plug-load asset exhibits an extreme, sustained spike during a weekend, the AI identifies potential electricity theft or equipment malfunction:

Possible explanation: The sustained consumption on a specific plug load suggests the unauthorized use of high-power external devices or potential electricity theft. Recommended action: Inspect the physical location of the asset for unauthorized hardware and configure a real-time rule-engine alert for future weekend consumption spikes on this circuit.

The finalized anomaly object, including the diagnostic summary, financial impact, and AI-generated guidance, is persisted in the central anomalies table for frontend visualization.

5.9. Tenant-Specific Configuration and Parameterization

To ensure the system remains adaptable to different operational requirements and economic conditions, the implementation includes a dedicated configuration layer for tenant-specific parameters. These settings are persisted in a centralized `anomaly_config` table and are refreshed dynamically by the `TenantRegistry` to govern the behavior of the `ADWorker` loops.

The configuration table allows for granular control over the detection sensitivity and the financial logic applied to each organization. This structure ensures that the system can be tailored to the specific risk tolerance and cost structures of diverse tenants. Key parameters include:

Sensitivity (q) This parameter defines the quantile at which an anomaly is triggered.

While the standard value is 0.99, a tenant may adjust this to increase or decrease the width of the normative band.

Check interval (minutes) This setting determines the frequency of the `TenantWorkerLoop` execution. While the default is 15 minutes, a tenant may increase this interval if they are primarily interested in higher-level temporal aggregations, such as daily reports, rather than real-time quarter-hourly monitoring.

Financial impact threshold This parameter governs the `AttributeDataProcessor` filtering logic. If set to a value such as 1.00, any meter whose peak daily consumption represents less than \$1.00 in potential waste is excluded from the monitoring and RCA processes to optimize computational resources.

Financial impact alert threshold This value serves as a final filter for the `ADPipeline` before an event is persisted. For example, if a tenant sets this to 10.00, the system

will only trigger a formal anomaly alert if the calculated financial impact exceeds \$10.00, thereby reducing alert fatigue caused by negligible deviations.

Energy cost per kWh Each tenant specifies their current electricity rate here. This value is the primary scalar used to convert energy residuals into the monetary impacts stored in the anomalies table.

Audit metadata Each configuration record includes `modified_by` and `modified_at` fields to maintain a traceable history of administrative changes to the detection parameters.

5.10. Frontend Visualization and User Interaction

The results of the backend orchestration and AI-driven diagnostics are presented through a multi-tenant frontend interface. This interface acts as the visualization layer for the anomalies persisted in the database, allowing facility managers to monitor energy health and interact with detected faults. The application layer utilizes a widget-based system to display real-time telemetry alongside anomaly indicators.

5.10.1. The Anomalies Table Interface

The implementation introduces an anomalies list as a core functional component within the platform's alert center. This module is designed to mirror the existing alarms list to ensure a consistent user experience while providing specialized tools for energy-centric data management. The interface is extended with an "Anomalies" tab located adjacent to the existing "Alarms" section, providing a unified environment for system health monitoring.

The data grid presents a structured overview of all identified deviations to facilitate rapid triage and operational decision-making. Each entry in the table displays critical diagnostic metrics, including the severity, the source asset, the financial impact, and the comparison between predicted and actual values. Columns for timeframes and tags enable the filtering and sorting of anomalies based on specific operational scopes.

5.10.2. User Feedback and Status Management

A central feature of the frontend is the capacity for human operators to validate the findings of the autonomous detection system. This human-in-the-loop interaction, in

The screenshot displays the Eliona test environment's anomalies list. The main area is a table with the following columns:

Severity	Source	Type	Financial Impact	Validity	Tags	Timeframe	Predicted	Actual	Deviation	Started	Ended	Status set by	Status set at	ID
Medium	Elektrozähler Lüftung Energie	VIRTUAL Elektrozähler	-49,97 €	confirmed	C. m. Proj.	0th 15m	6'634.728 kWh	6'658.39 kWh	+223.663 kWh	16.12.2025, 12:45	16.12.2025, 12:30	bjoern.erb@eliona.io	24.12.2025, 13:31	2107
Medium	Elektrozähler Lüftung Energie	VIRTUAL Elektrozähler	-66,07 €	not set	C. m. Proj.	0th 15m	6'034.762 kWh	6'316.63 kWh	+281.868 kWh	16.12.2025, 12:15	16.12.2025, 12:30		16.12.2025, 12:15	2106
Medium	Elektrozähler Lüftung Energie	VIRTUAL Elektrozähler	-69,99 €	confirmed	C. m. Proj.	0th 15m	5'472.15 kWh	5'759.31 kWh	+286.16 kWh	16.12.2025, 12:00	16.12.2025, 12:15	bjoern.erb@eliona.io	24.12.2025, 13:31	2105
Medium	ESG Report outputs_ virtual Heating,Total	VIRTUAL ESG-Report	+40,49 €	not set	Cont... mediu...	0th 15m	308.326 kWh	154.35 kWh	-153.976 kWh	16.12.2025, 12:00	16.12.2025, 12:15		16.12.2025, 12:00	2103
Medium	Electricity total	VIRTUAL ESG-Report	+93,86 €	confirmed	Cont... mediu...	0th 15m	752.482 kWh	385.78 kWh	-366.702 kWh	16.12.2025, 12:00	16.12.2025, 12:15	bjoern.erb@eliona.io	24.12.2025, 13:31	2104
Medium	Elektrozähler Lüftung Energie	VIRTUAL Elektrozähler	-71,61 €	false	C. m. Proj.	0th 15m	4'901.482 kWh	5'189.47 kWh	+287.988 kWh	16.12.2025, 11:45	16.12.2025, 12:00	bjoern.erb@eliona.io	24.12.2025, 13:31	2102
Medium	Elektrozähler Klima_001_OX Energie	VIRTUAL Elektrozähler	-45,59 €	false	C. m. Proj.	0th 15m	7'851.011 kWh	8'052.64 kWh	+201.629 kWh	16.12.2025, 11:30	16.12.2025, 11:45	bjoern.erb@eliona.io	24.12.2025, 13:31	2101
Medium	Elektrozähler WP Energie	VIRTUAL Elektrozähler	-61,54 €	not set	C. m. Proj.	0th 15m	87.251 kWh	326.35 kWh	+239.099 kWh	16.12.2025, 11:15	16.12.2025, 11:30		16.12.2025, 11:15	2098
Medium	Elektrozähler Klima_001_OX Energie	VIRTUAL Elektrozähler	-45,09 €	not set	C. m. Proj.	0th 15m	8'425.175 kWh	8'613.78 kWh	+188.605 kWh	16.12.2025, 11:15	16.12.2025, 11:30		16.12.2025, 11:15	2100
Low	ESG Report plausibilität Star Renewable Electricity	ESG-Report - input - scripts	+1,26 €	not set	low	0th 15m	39.983 kWh	34.93 kWh	-5.053 kWh	16.12.2025, 11:15	16.12.2025, 11:30		16.12.2025, 11:15	2099
Medium	Elektrozähler Klima_001_OX Energie	VIRTUAL Elektrozähler	-44,79 €	not set	C. m. Proj.	0th 15m	8'886.674 kWh	9'093.1 kWh	+206.426 kWh	16.12.2025, 11:00	16.12.2025, 11:15		16.12.2025, 11:00	2097
Medium	Elektrozähler WP Energie	VIRTUAL Elektrozähler	-62,39 €	not set	C. m. Proj.	0th 15m	85.508 kWh	327.56 kWh	+242.052 kWh	16.12.2025, 11:00	16.12.2025, 11:15		16.12.2025, 11:00	2095
Low	ESG Report plausibilität Star Renewable Electricity	ESG-Report - input - scripts	+1,06 €	not set	low	0th 15m	44.583 kWh	40.31 kWh	-4.273 kWh	16.12.2025, 11:00	16.12.2025, 11:15		16.12.2025, 11:00	2096
Medium	Elektrozähler Unterverteiler	VIRTUAL Elektrozähler	-1,06 €	not set	medium	0th 15m	1.060.000 kWh	1.060.000 kWh	0 kWh	16.12.2025, 11:00	16.12.2025, 11:15		16.12.2025, 11:00	2094

Figure 5.2: Anomalies list interface in the Eliona frontend, showing key diagnostic metrics, financial impact, and filtering options.

which human feedback is integrated into an automated process, is required to maintain model accuracy and long-term system trust.

Operators utilize a dynamic status control to select one or multiple anomalies and assign a status of “Not set”, “Confirmed”, or “False positive”. A false positive is defined as a result that indicates a condition is present when it is not. The “Add comment” function allows facility managers to record observations or remediation steps directly within the anomaly record, creating a traceable history for maintenance teams.

Marking an anomaly as a false positive triggers a data-cleansing process. This information is utilized to exclude the flagged period from future training cycles. As a result, non-anomalous events, such as authorized maintenance or unique operational shifts, are correctly classified as normal behaviour in future baselines.

5.10.3. Integrated Anomaly Analytics and Visualization

The implementation extends the platform’s analytical capabilities by introducing a specialized anomaly-detection analytic type. This component integrates directly into the existing insight analytics framework, allowing users to visualize stochastic detection results across dashboards and automated reports. It is designed to provide a cohesive visual bridge between raw telemetry and the probabilistic outputs of the Chronos-2 model.

Dynamic Chart Overlays

When the anomaly-detection analytic is active, the platform provides two distinct layers of visual context to the consumption charts. First, the system generates a red background overlay for the entire asset profile whenever a deviation is identified. This highlight persists across all associated attributes of the asset, regardless of the selected aggregation level, enabling operators to observe how a failure in one sub-meter propagates through the energy consumption of the larger system.

Second, for specific energy meters, the interface provides a detailed granular view when the timeframe and aggregation type match the detection parameters. It displays red dots to mark individual anomalous data points alongside a green expected range. This green band represents the stochastic normative profile established by the model's quantiles, allowing users to visually quantify the magnitude of the deviation relative to the predicted operational baseline.

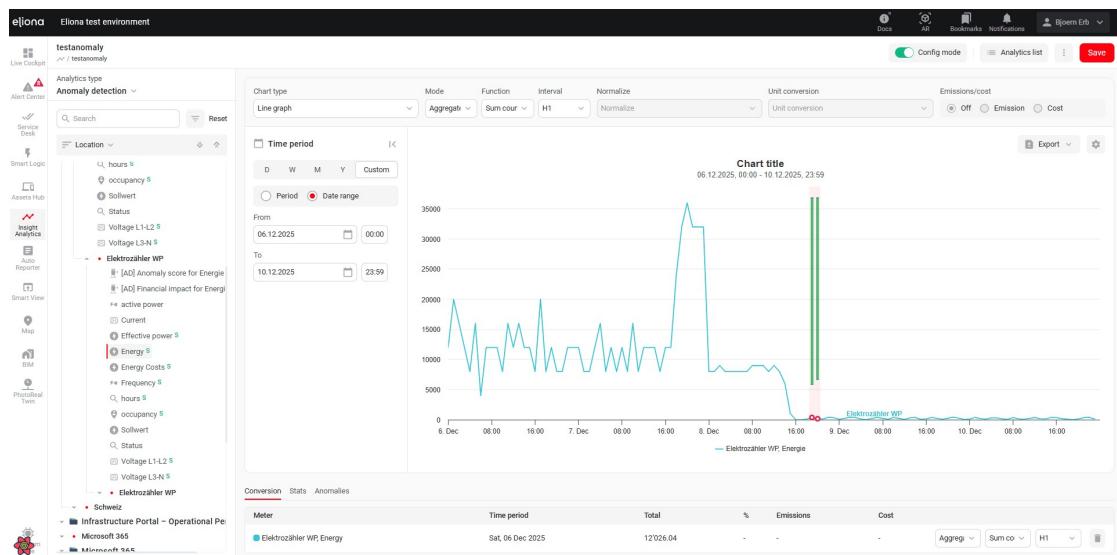


Figure 5.3: Integrated anomaly-detection analytic within the insight analytics framework, highlighting anomalous periods on the consumption charts.

Interactive Diagnostics and Tooltips

To facilitate immediate root-cause identification, the system includes interactive tooltips that appear upon hovering over an anomalous data point. These tooltips provide a concentrated summary of the anomaly object, including:

Quantitative metrics Real-time data regarding the financial impact, the predicted versus actual consumption, and the severity level.

AI-synthesized context The AI explanation regarding the likely origin of the fault (for example, “extended equipment runtime or a control issue”) and the corresponding

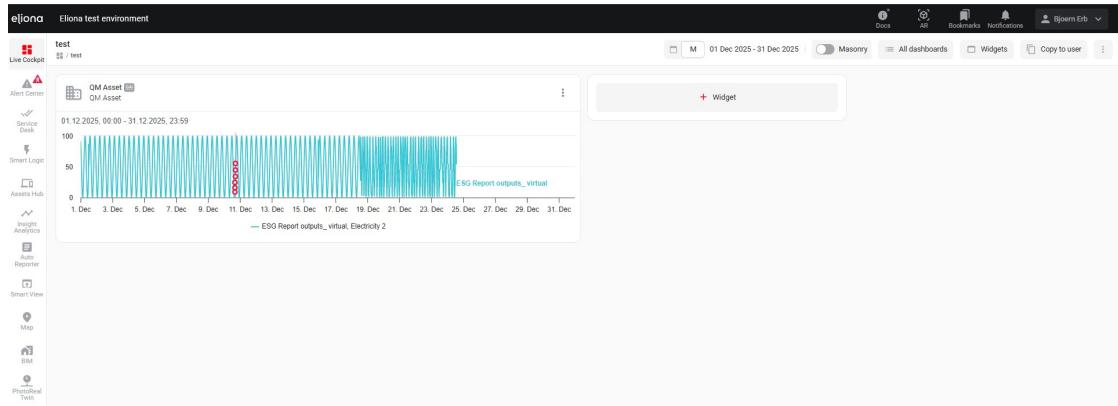


Figure 5.4: Anomaly-detection analytic embedded in a customizable dashboard, combining stochastic overlays with other operational widgets.

recommended action (for example, “review the affected equipment settings and operating schedule”).

Validation status The current validity status (for example, “Confirmed”), which reflects the user feedback provided in the anomalies list.

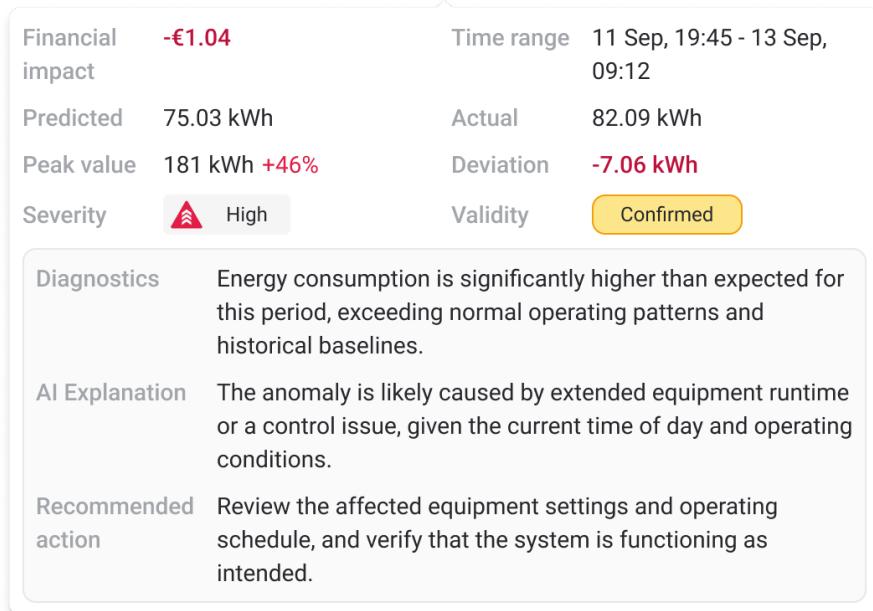


Figure 5.5: Interactive tooltip attached to an anomalous data point, summarizing financial impact, AI explanation, and validation status.

Because this functionality is implemented as a standard analytic type, it is fully compatible with the platform’s existing reporting engine and customizable dashboards. This ensures that the detection results are not isolated within a separate module but are

integrated into the daily operational workflows of the facility management team.

5.10.4. Anomaly Detail View and Operational Synthesis

The anomaly detail view provides a specialized environment for the deep-dive analysis of individual detection events. Users access this page by selecting the navigational arrow corresponding to a specific entry in the anomalies list. This view consolidates all relevant quantitative and qualitative data into a single operational summary, facilitating a transition from detection to remediation.

The detail page is structured to provide immediate clarity regarding the severity and context of the fault. The header section displays the core metadata, including the financial impact, the severity level, and the validation status. A dedicated validation panel allows the user to update the status to “Confirmed” or “False” and provides a text area for recording technical comments or observations.

The primary analytical component of this view is a high-resolution analytic chart. This visualization is automatically centered on the anomalous period, allowing the user to observe the consumption profile immediately before and after the identified deviation. This spatial centering provides critical temporal context, helping the operator determine whether the event was a singular spike or the onset of a sustained operational shift.

The lower section of the page presents the diagnostics and AI-synthesis results. The diagnostics string summarizes the technical conditions, such as the outside temperature and weather at the time of the event, and identifies the likely contributing sub-meters. Below this, the AI explanation provides a natural-language interpretation of the data, for instance by identifying a likely “data acquisition failure” or “unexpected shutdown of major loads”. The recommended actions then list specific investigative steps, such as checking metering infrastructure for outages or verifying scheduled maintenance activities.

By combining these interactive elements, the anomaly detail view transforms raw telemetry into a structured maintenance task, directly supporting the actionable-insights requirement established in the system’s methodology.

5.10.5. Anomaly Statistics and Macro-Level Reporting

To complement the detailed, asset-specific views, the platform includes an anomaly-statistics dashboard designed for macro-level analysis and executive reporting. This interface aggregates data across entire tenants or individual sites, providing facility managers with a high-level overview of system health and financial performance over

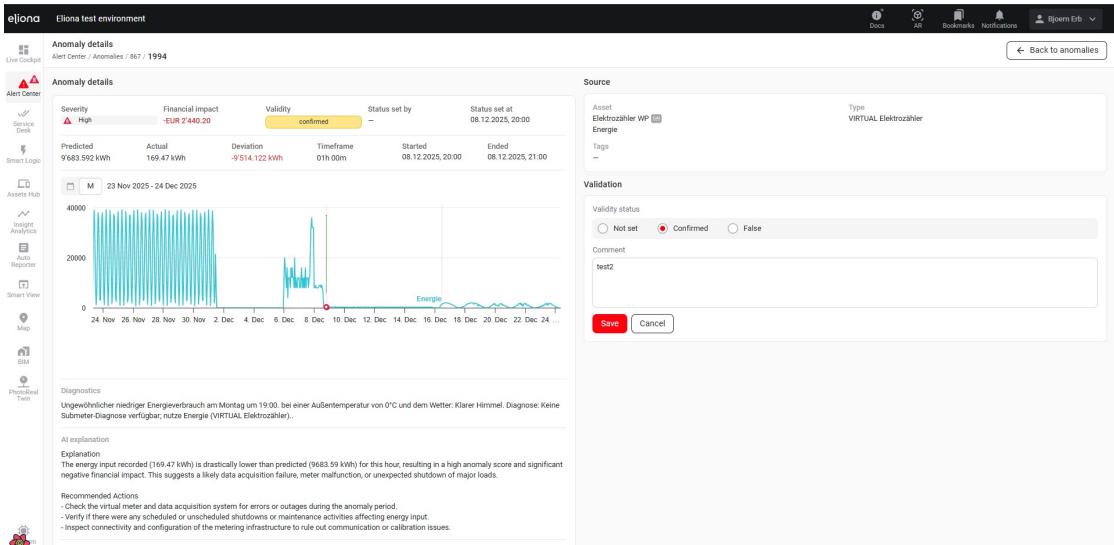


Figure 5.6: Anomaly detail view aggregating financial impact, validation status, analytic chart, diagnostics, and AI-synthesized recommended actions.

time.

The scope of the data presented is controlled through a unified timeframe selector. Users can toggle between standard periods, such as the current month or the current fiscal year, which instantly recalculates all metrics on the dashboard. The primary key performance indicators (KPIs) focus on the aggregated financial impact. A prominent widget displays the total monetary value of detected energy waste or savings within the selected period, alongside a percentage-based comparison to the previous equivalent timeframe. This immediate context allows stakeholders to quickly determine if operational performance is trending positively or negatively.

Beneath the top-level KPIs, the dashboard provides categorical breakdowns to identify systemic issues. A visualization of the most impact by asset type highlights which categories of equipment, such as HVAC or lighting, are contributing most significantly to financial losses. Temporal trends are visualized through an overspend-versus-saved bar chart, allowing users to visually track the effectiveness of remediation efforts over successive periods. A parallel bar chart provides a count of anomalies per asset type, distinguishing between high-frequency, low-impact events and rare, high-cost failures. For multi-site tenants, a site-aggregation list details the anomaly count and total impact for each location. Selecting a specific site in this list redirects the user to a filtered version of the statistics page dedicated solely to that location's data.

A critical analytical tool within this dashboard is the financial-impact heatmap, designed to reveal temporal patterns in energy waste. The configurations of its axes adapt dynamically based on the selected timeframe. For shorter periods, the x-axis represents the date and the y-axis represents the time of day. The cell colour intensity

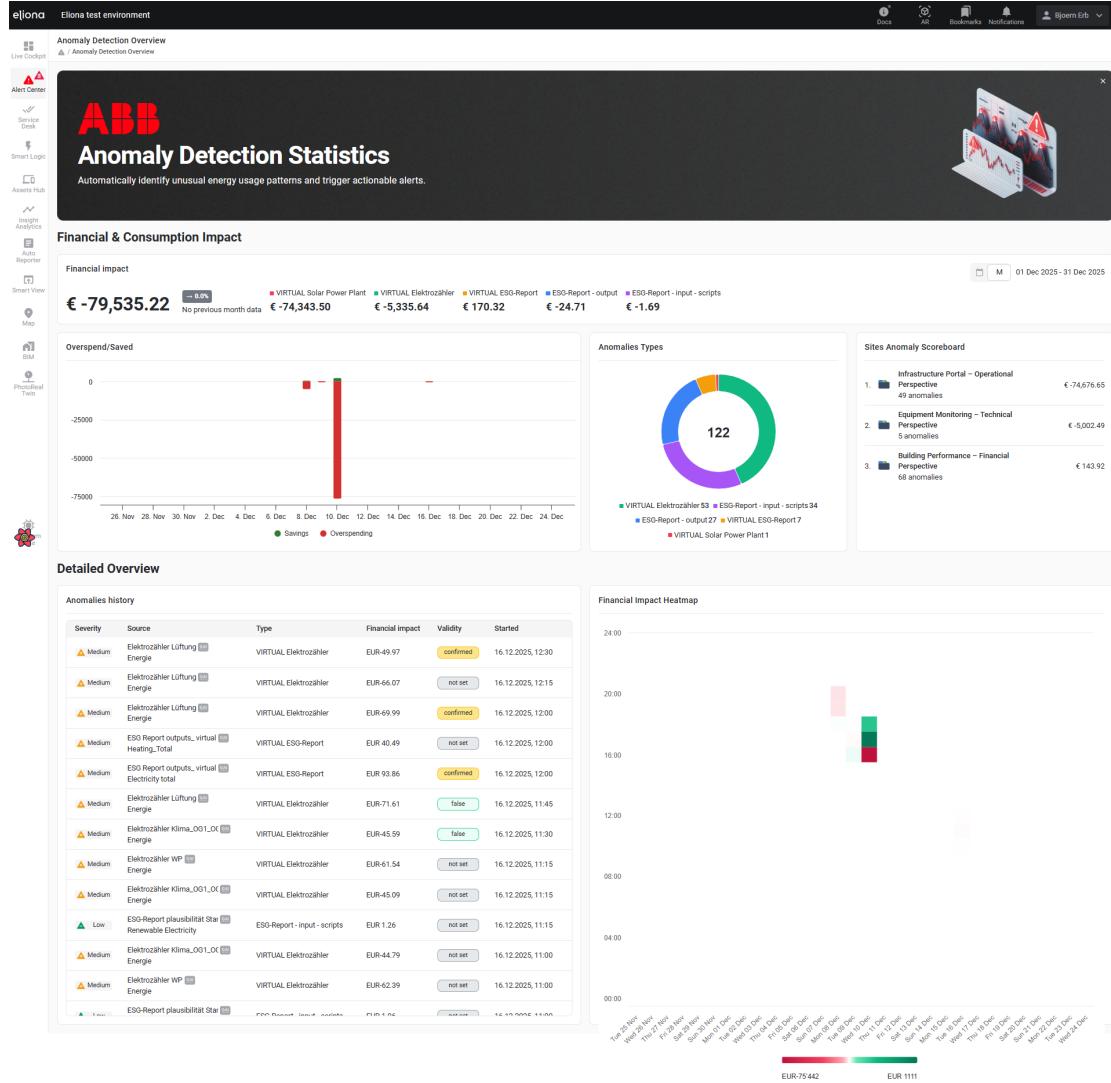


Figure 5.7: Anomaly-statistics dashboard providing macro-level KPIs, categorical breakdowns, temporal trends, and site-level aggregation for executive reporting.

indicates the magnitude of financial impact, making it immediately apparent if specific times of day consistently experience costly anomalies. For longer timeframes, such as a full year, the heatmap aggregates data to show day of week versus time of day. This view is particularly valuable for identifying systemic operational faults, such as recurring high-impact anomalies on Monday mornings that may indicate faulty equipment start-up sequences after weekend shutdowns.

5.10.6. Asset-Specific Anomaly Integration

The platform's asset-detail view is expanded to include a dedicated "Anomalies" tab, facilitating a seamless transition between general asset monitoring and specialized anomaly investigation. This view is filtered to display only the deviations and financial impacts associated with the meters assigned to the currently selected asset.

The interface provides a localized summary that includes the date of the most recent anomaly and the total financial impact accumulated by the asset's meters within the selected timeframe. A donut chart visualizes the distribution of anomaly types, while a localized "Anomalies on chart" widget highlights the specific timestamps where deviations occurred relative to the asset's consumption profile. An "Anomalies history" table lists the individual events, including their severity and financial impact, providing a direct link to the anomaly detail view for further investigation.

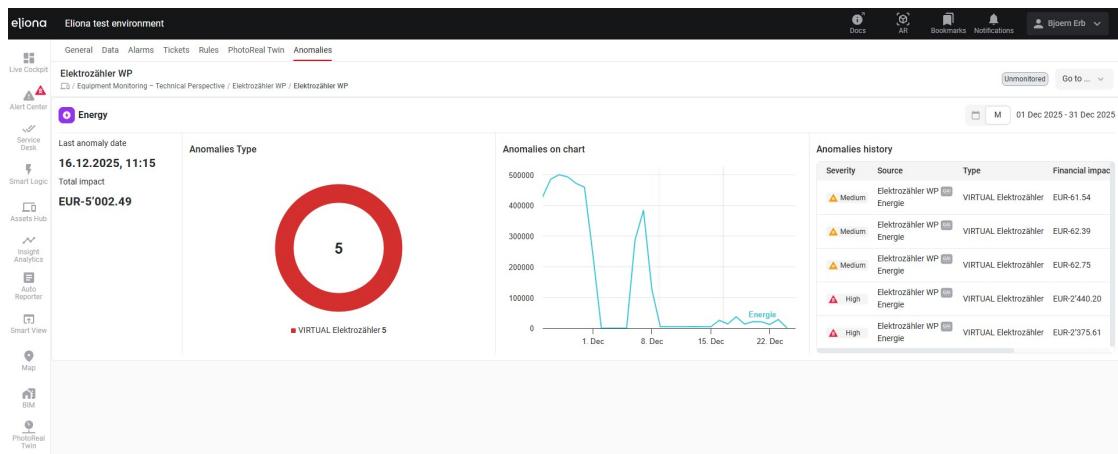


Figure 5.8: Asset-detail view with integrated anomalies tab, showing localized statistics, anomaly-type distribution, and historical events for the selected asset.

6

Discussion and Future Work

The evaluation of the integrated system demonstrates the efficacy of combining stochastic forecasting with hierarchical root-cause analysis. However, the transition from a synthetic benchmark to diverse industrial environments reveals specific areas where the methodology can be refined to enhance diagnostic depth and modelling flexibility.

6.1. Critical Reflection on System Design

The current implementation of root-cause analysis (RCA)—the process of identifying the origin of a fault—relies primarily on the statistical attribution of financial impact across sub-meters. While this identifies *where* an anomaly occurs, it does not fully explain *why* the deviation was triggered within the control layer. Future iterations could enhance the RCA by integrating the operational states of high-consumption assets. By analysing control signals, such as valve positions or modulation frequencies, the system could verify whether a consumption spike is a legitimate response to a manual override or a failure of the underlying control logic.

Furthermore, the feature set utilized in the ADPipeline is currently restricted to meteorological data and temporal indicators. Empirical evidence from the BOPTEST correlation analysis (see Figure 4.12) indicates that occupancy—represented by people count—is one of the most significant drivers of energy consumption. The absence of real-time occupancy telemetry in many tenant buildings represents a significant information gap. The system should be expanded to allow tenants to select specific attributes, such as CO₂ levels or access-control logs, to be included as exogenous

features for customized detection models.

6.2. Data Integrity and User-Centric Baseline Selection

The current modelling approach assumes that all historical data preceding the activation of an anomaly-detection licence represents a healthy operational state. This assumption is often violated in real-world facilities where persistent faults may already be present. To address this, an interface for manual baseline configuration is proposed. This would enable facility managers to designate specific historical intervals as *gold-standard* periods.

Implementing this with the current Chronos-2 architecture presents a technical challenge, as the model requires a continuous temporal sequence immediately preceding the target timestamp. To utilize a non-contiguous healthy baseline from a previous year, the system would require a mechanism to translate and align historical timestamps to the current prediction window.

Additionally, the reliability of the RCA is heavily dependent on the metering density of the building. In facilities with low sub-metering granularity, the system's ability to attribute faults remains limited to large-scale aggregates, highlighting the need for further evaluation on diverse, real-world datasets.

6.3. Future Architecture: The Universal Energy Feature Forecaster

The utilization of Chronos-2 in the current system represents an adaptation of a sequential foundation model for feature-based anomaly detection. While effective, this approach does not fully leverage the model's internal probability distributions for anomaly scoring in the same way a mixture-density network (MDN) does. A significant limitation is the reliance on fixed quantile bounds (for example, $q_{0.99}$), which simplifies the complex multimodal output of the transformer into a binary threshold.

6.3.1. In-Context Zero-Shot Modelling

A proposed architectural shift involves the development of a specialized foundation model designed as a universal energy feature forecaster. Instead of predicting a sequence based on recent history, this model would utilize in-context learning (ICL). In

this paradigm, the model is provided with a set of baseline features and their corresponding target values as context, regardless of their temporal proximity to the current timestamp. This would allow the model to ingest healthy baseline data from a different season or a different year as a direct reference for the current prediction task.

6.3.2. Probabilistic Anomaly Scoring

The proposed model would retain the token-based probability output of the Chronos architecture but utilize the full distribution to calculate an anomaly score. By computing the negative log-likelihood (NLL) of an observed value across all predicted tokens, the system would achieve a sensitivity comparable to the MDN while maintaining the zero-shot generalization of a foundation model.

This architecture would also enable comparative baseline analysis without retraining. For example, an operator could predict March consumption using both January and February baselines to quantify the resulting energy savings from efficiency measures implemented in February. In doing so, the anomaly-detection system would evolve from a pure fault detector into a broader decision-support tool for evaluating building-decarbonization strategies.

6.4. Reflections on Energy Anomaly Benchmarking

The experimental evaluation conducted within this research represents a foundational step towards a standardized benchmarking framework for energy-specific anomaly detection. It highlights the necessity for datasets that prioritize **contextual anomalies**—deviations that are only anomalous relative to external variables such as weather or occupancy—over simple point deviations. By utilizing the **volume under the surface of the precision-recall curve (VUS-PR)**, the benchmark addresses the inherent temporal characteristics of industrial energy faults, which frequently persist over extended durations.

However, the current benchmarking methodology reveals several opportunities for improvement. While the results provide a comparative overview, the absence of exhaustive **hyperparameter tuning**—the process of optimizing the internal parameters of a model to achieve peak performance—for many baseline methods potentially masks their true detection capabilities. Furthermore, the protocol for comparing zero-shot **foundation models**—large-scale models pre-trained on diverse datasets to perform tasks without site-specific training—against traditional supervised methods requires further formalization.

Future benchmarking efforts must ensure that all models are evaluated under optimal configurations to support a scientifically robust comparison. Such a refined framework will serve as a vital tool for the objective validation of emerging architectures in the building-automation sector.

7

Conclusion

The research conducted in this thesis successfully established a robust methodology and technical framework for the automated detection and quantification of energy anomalies in building environments. By addressing the statistical complexities of building telemetry—specifically its non-stationary and multi-modal nature—the project provided a solution that transcends the limitations of traditional deterministic forecasting.

The investigation into predictive modelling paradigms revealed that standard sequential forecasting is susceptible to error propagation and the adaptation paradox. It was shown that autoregressive models often incorporate anomalous data into their internal state, which leads to signal instability and the masking of sustained faults. To resolve these issues, a stochastic approach was developed that utilizes the **Chronos-2** foundation model within a feature-driven inference strategy. This methodology establishes a probabilistic normative operational band, allowing for the reliable identification of **Multivariate Context Point Anomalies (MCPA)**—deviations that are only identifiable through the joint analysis of consumption and exogenous drivers such as weather and occupancy.

The technical realization was achieved through a distributed, polyglot architecture. A high-performance **Scala** microservice managed multi-tenant isolation and data orchestration, while a **Python** endpoint hosted on **Azure Machine Learning** provided the required predictive capacity. The integration of a hierarchical **Root Cause Analysis (RCA)** and **Generative AI** synthesis transformed raw detection results into actionable operational intelligence. It was demonstrated that by attributing financial impact to specific assets and generating natural-language remediation steps, the system provides facility managers with a transparent tool for energy waste mitigation.

The implementation of specialized frontend modules—including the anomalies list,

status management, and temporal heatmaps—ensures that detection events are integrated into standard maintenance workflows. The capacity for human-in-the-loop feedback, where users validate anomalies as confirmed or false, establishes a mechanism for continuous data cleansing and model refinement. This ensures that the system remains accurate as building characteristics evolve over time.

Ultimately, this research moved the **Eliona** platform from a state of reactive alarm management to proactive, intelligence-driven energy monitoring. The shift from point-based detection to stochastic distribution modeling enables a precise calculation of financial residuals, providing organizations with quantifiable data to support their sustainability and decarbonization objectives. While future work remains regarding the inclusion of control-layer states and the development of in-context forecasters, this project provides a scientifically validated foundation for the next generation of energy management systems in the building-automation sector.

References

- [Rot+04] Kurt W Roth et al. “The energy impact of faults in US commercial buildings”. In: (2004).
- [Ant09] Pedro Antmann. “Reducing technical and non-technical losses in the power sector”. In: (2009).
- [MM09] Patrick McDaniel and Stephen McLaughlin. “Security and privacy challenges in the smart grid”. In: *IEEE security & privacy* 7.3 (2009), pp. 75–77.
- [Su+19] Ya Su et al. “Robust anomaly detection for multivariate time series through stochastic recurrent neural network”. In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 2828–2837.
- [Pap+22] John Paparrizos et al. “TSB-UAD: an end-to-end benchmark suite for univariate time-series anomaly detection”. In: *Proceedings of the VLDB Endowment* 15.8 (2022), pp. 1697–1711.
- [Alš24] Oskaras Alšauskas. “World energy outlook 2024”. In: *International Energy Agency: Paris, France* (2024).
- [Edi24] Edison Foundation Institute for Electric Innovation. *120 million smart meters in US in 2022*. Online article. 2024. URL: <https://www.enlit.world/library/120-million-smart-meters-in-us-in-2022> (visited on 12/24/2025).
- [IoT24] IoT Analytics. *Global Smart Electricity Meter Adoption 2024 by Region*. Online graphic. 2024. URL: <https://iot-analytics.com/wp-content/uploads/2024/02/Global-Smart-Electricity-Meter-Adoption-2024-by-Region-vweb.png> (visited on 12/25/2025).
- [LP24] Qinghua Liu and John Paparrizos. “The elephant in the room: Towards a reliable time-series anomaly detection benchmark”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 108231–108261.
- [Ans+25] Abdul Fatir Ansari et al. “Chronos-2: From univariate to universal forecasting”. In: *arXiv preprint arXiv:2510.15821* (2025).

- [Azz+25] Davide Azzalini et al. “An empirical evaluation of deep autoencoders for anomaly detection in the electricity consumption of buildings”. In: *Energy and Buildings* 327 (2025), p. 115069. ISSN: 0378-7788. DOI: <https://doi.org/10.1016/j.enbuild.2024.115069>. URL: <https://www.sciencedirect.com/science/article/pii/S037877882401185X>.
- [Eli25a] Eliona IoT Platform. *Asset Modeling – Create Templates*. Online documentation. 2025. URL: <https://doc.eliona.io/collection/eliona-english/documentation/assets/asset-modeling-create-templates> (visited on 12/23/2025).
- [Eli25b] Eliona IoT Platform. *Assets*. Online documentation. 2025. URL: <https://doc.eliona.io/collection/eliona-english/documentation/assets> (visited on 12/23/2025).
- [Eli25c] Eliona IoT Platform. *Introduction to Ontologies*. Online documentation. 2025. URL: <https://doc.eliona.io/collection/eliona-english/academy/introduction-to-ontologies> (visited on 12/20/2025).
- [Eli25d] Eliona IoT Platform. *Rule Chains*. Online documentation. 2025. URL: <https://doc.eliona.io/collection/eliona-english/documentation/alarms-rules-and-escalations/rule-chains> (visited on 12/23/2025).
- [Eli25e] Eliona IoT Platform. *Rules*. Online documentation. 2025. URL: <https://doc.eliona.io/collection/eliona-english/documentation/alarms-rules-and-escalations/rules> (visited on 12/23/2025).
- [Eli25f] Eliona IoT Platform. *Structuring Assets*. Online documentation. 2025. URL: <https://doc.eliona.io/collection/eliona-english/documentation/assets/structuring-assets> (visited on 12/23/2025).
- [HHA25] Basu Hela, Praveen Prasad Handigol, and Pandarasamy Arjunan. “Are Time Series Foundation models good for Energy Anomaly Detection?” In: *Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems*. E-Energy ’25. Association for Computing Machinery, 2025, pp. 656–665. ISBN: 9798400711251. DOI: [10.1145/3679240.3734633](https://doi.org/10.1145/3679240.3734633). URL: <https://doi.org/10.1145/3679240.3734633>.
- [MM25] Roya Morshedi and S. Mojtaba Matinkhah. “A Comprehensive Review of Deep Learning Techniques for Anomaly Detection in IoT Networks: Methods, Challenges, and Datasets”. In: *Engineering Reports* 7.9 (2025), e70415. DOI: <https://doi.org/10.1002/eng2.70415>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/eng2.70415>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/eng2.70415>.