Hochschule Konstanz
Department of Computer Science

Submitted by
Samuel Tim
Student Number 307636

samuel.tim200yahoo.de

Bachelor Thesis

Energy Anomaly Detection with Machine Learning

Konstanz, 31st December 2025

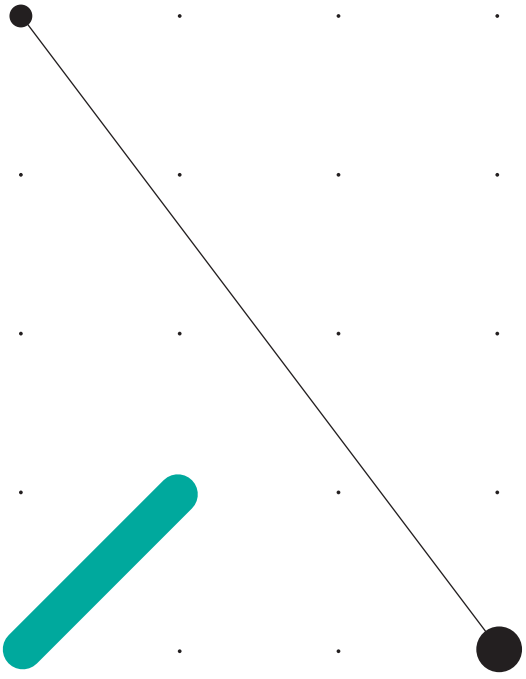
Hochschule Konstanz
Department of Computer Science

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
2.2 Foundations of Anomaly Detection . . . 12
2.2.1 Dimensionality and Normality Regimes . . . 12
2.2.2 Terminology: Multivariate and Multi-Target Time Series . . . 13
2.2.3 Structural Classes of Anomalies . . . 13
2.2.4 Multiplicity of Occurrence . . . 14
2.3 Methodological Approaches to Anomaly Detection . . . 14
2.3.1 Anomaly Scores . . . 15
2.3.2 Learning Paradigms . . . 15
2.3.3 Families of Detection Methods . . . 15
2.4 Benchmarking Foundations . . . 17
2.4.1 Binary Labels and Confusion Matrix . . . 17
2.4.2 Evaluation Metrics . . . 17
2.5 Synthesis of Foundations . . . 17

3 Related Work **19**
3.1 Classical Energy Baseline and Rule-Based Detection . . . 19
3.2 Reliability and Benchmarking: The TSB-AD Framework . . . 20
3.2.1 Systemic Flaws and Metric Reliability . . . 20
3.2.2 Benchmark Evaluation and Model Hierarchy . . . 20
3.2.3 Implications for Multivariate Context Point Anomalies . . . 21
3.2.4 Large-Scale Supervised Energy Benchmarks: LEAD 1.0 . . . 22
3.3 Comparative Analysis of Deep Learning and Foundation Models in Energy Systems . . . 23
3.3.1 Deep Generative Models and the Advantage of Reconstruction . . . 23
3.3.2 Time-Series Foundation Models in the Energy Domain . . . 23
3.3.3 Synthesis of Related Work . . . 24

4 Methodology **25**
4.1 System Context: The Eliona IoT Platform . . . 25
4.1.1 Modular System Architecture . . . 25
4.1.2 Asset Modeling and Hierarchical Ontology . . . 26
4.2 Formal Design Objectives and System Requirements . . . 27
4.2.1 Functional Objectives . . . 27
4.2.2 Operational Objectives . . . 27
4.3 Financial Impact Quantification . . . 28
4.3.1 Distribution-Aware Baseline . . . 28
4.3.2 Fallback Without Mixture Information . . . 28
4.3.3 Economic Impact . . . 29
4.4 Hierarchical Root Cause Analysis and Action Synthesis . . . 29
4.4.1 Ontology-Guided Attribution . . . 29
4.4.2 Aggregation by Asset Type . . . 29
4.4.3 Contextual Synthesis and Action Generation . . . 30
4.4.4 Design Rationale . . . 30
4.5 Critique of Sequential Forecasting for Anomaly Detection . . . 30
4.5.1 Synthetic Experimental Setup . . . 30
4.5.2 Failure Mode 1: Error Propagation and Instability . . . 31
4.5.3 Failure Mode 2: Rapid Adaptation and the PA-F1 Illusion . . . 32
4.5.4 Mitigation Strategies . . . 33
4.6 Statistical Limitations of Point and Gaussian Predictions . . . 34
4.6.1 The Failure of Mean Squared Error Minimization . . . 34
4.6.2 The Gaussian Distribution Paradox . . . 35
4.6.3 Solution: Mixture Density Networks . . . 36

4.7 Distribution-Aware Anomaly Scoring for Mixture Density Models 38
4.7.1 Mean Residual: Failure Under Multimodality 38
4.7.2 Probability Integral Transform (PIT) 38
4.7.3 Negative Log-Likelihood and Its Limitations 40
4.7.4 Density–Quantile (DQ) Probability 40
4.7.5 Density–Quantile Severity Scaling 41
4.7.6 Summary 41
4.8 Methodological Scope 41

5 Benchmarking **43**
5.1 Benchmark Design and Dataset Generation 43
5.2 Feature and Target Definition 43
5.3 Data Segmentation and Anomaly Injection 44
5.4 Evaluation Constraints and Benchmark Limitations 45
5.4.1 Training Stability and Coverage Bias 45
5.4.2 Comparability Across Model Classes 46
5.4.3 Interpretation Scope 46
5.5 Comparative Model Performance and Structural Evaluation 46
5.5.1 Stochastic and Hybrid Architectures 46
5.5.2 Training Stability and Classical Baselines 48
5.5.3 Season-Matched Three-Month Evaluation 48
5.5.4 Seasonal Translation Sensitivity 49
5.5.5 Model Selection Rationale 49

6 Implementation **51**
6.1 Integrated System Architecture and Technology Stack 51
6.1.1 Deployment and Cluster Integration 51
6.1.2 Data Orchestration and Persistence 52
6.2 Chronos-2 Analytics Endpoint 52
6.3 Scala Microservice: Multi-Tenant Orchestration 53
6.3.1 Multi-Tenant Lifecycle Management 53
6.4 Data Acquisition and Processing Pipeline 53
6.4.1 Attribute Selection and Hierarchical Scoping 54
6.4.2 Contextual Enrichment 54
6.4.3 Data Conditioning and Gap Resilience 54
6.4.4 Reactive Synchronization 54
6.5 Stochastic Inference and Anomaly Quantification 54
6.5.1 Feature-Driven Inference 55
6.5.2 Quantile-Based Scoring 55

6.6 Hierarchical Root Cause Analysis (RCA) . . . 55
6.6.1 Diagnostic Attribution . . . 55
6.6.2 Localization and Contextual Enrichment . . . 55
6.7 Temporal Collapse and Persistence . . . 56
6.8 AI Synthesis and Action Recommendation . . . 56
6.9 Tenant-Specific Configuration . . . 56
6.10 Frontend Integration for Operational Decision Support . . . 56
6.10.1 Centralized Anomaly Registry and Validation Loop . . . 57
6.10.2 Quantile Visualization and Diagnostic Context . . . 57
6.10.3 Contextual Tooltips and AI Synthesis . . . 57
6.10.4 Anomaly Detail View and Action Synthesis . . . 59
6.10.5 Macro-Level Reporting and Portfolio Analytics . . . 59
6.10.6 Asset-Scoped Anomaly Integration . . . 59
6.11 Implementation Summary . . . 60

7 Discussion and Future Work **63**
7.1 Critical Reflection on System Design . . . 63
7.2 Data Integrity and User-Centric Baseline Selection . . . 64
7.3 Future Architecture: The Universal Energy Feature Forecaster . . . 64
7.3.1 In-Context Zero-Shot Modelling . . . 64
7.3.2 Probabilistic Anomaly Scoring . . . 65
7.4 Reflections on Energy Anomaly Benchmarking . . . 65

8 Conclusion **67**

A Additional Figures **69**
A.1 Training History Across All Meters . . . 69

References **71**

Glossary

anomaly observation or pattern that deviates significantly from a defined notion of normality. viii

benchmark standardized dataset and evaluation protocol used to compare the performance of different anomaly detection methods. viii

confusion matrix tabular summary of prediction results that counts true positives, true negatives, false positives, and false negatives. viii

ground truth reference labels that indicate for each observation whether it is considered normal or anomalous, used as a standard when evaluating detection performance. viii

mislabeling inconsistent assignment of anomaly labels to similar or identical patterns, which distorts evaluation by inflating false-negative rates. viii, 20

precision for anomaly detection, the proportion of predicted anomalous points or segments that are actually anomalous (true positives divided by all positive predictions). viii

recall for anomaly detection, the proportion of truly anomalous points or segments that are correctly detected (true positives divided by all actual anomalies). viii

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. viii, 20

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. viii, 20

Acronyms

AMI Advanced Metering Infrastructure. 1

BAS Building Automation Systems. 1

BOPTEST Building Optimization Performance Test Framework. 5

CNN convolutional neural network. 20, 21

IPMVP International Performance Measurement and Verification Protocol. 5

MCPA Multivariate Context Point Anomaly. 4

ML machine learning. 21

OmniAnomaly stochastic recurrent neural network model OmniAnomaly. 20

PA-F1 Point-Adjustment F1 score. 20

Sub-PCA subspace principal component analysis. 20

TSAD time series anomaly detection. 20

TSB-AD Time Series Benchmark for Anomaly Detection. 20, 21

TSFM time-series foundation model. 3

VUS-PR Volume Under the Surface–Precision Recall. 20, 21

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.

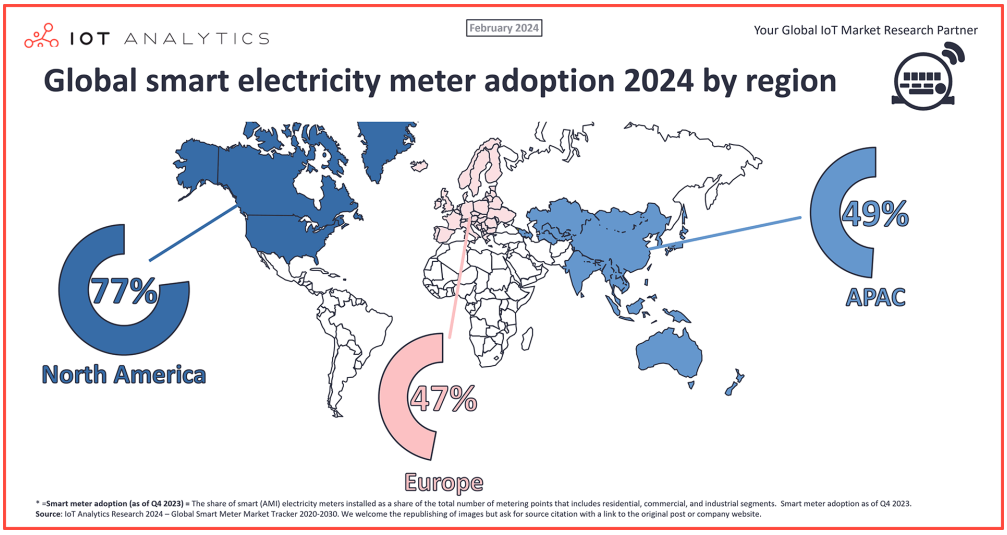


Figure 1.1: Global smart electricity meter adoption by region in 2024, illustrating the varying levels of AMI penetration across markets [IoT24].

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 5.1).
- 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 formal foundations required for contextual anomaly detection in building-energy telemetry. It characterizes the structural, statistical, and causal properties of building-energy data, defines the relevant anomaly taxonomies, and introduces the methodological and benchmarking concepts used throughout this thesis.

Based on these foundations, the chapter derives formal modeling requirements that constrain the design of detection, quantification, and attribution methodologies developed in the subsequent chapters.

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.

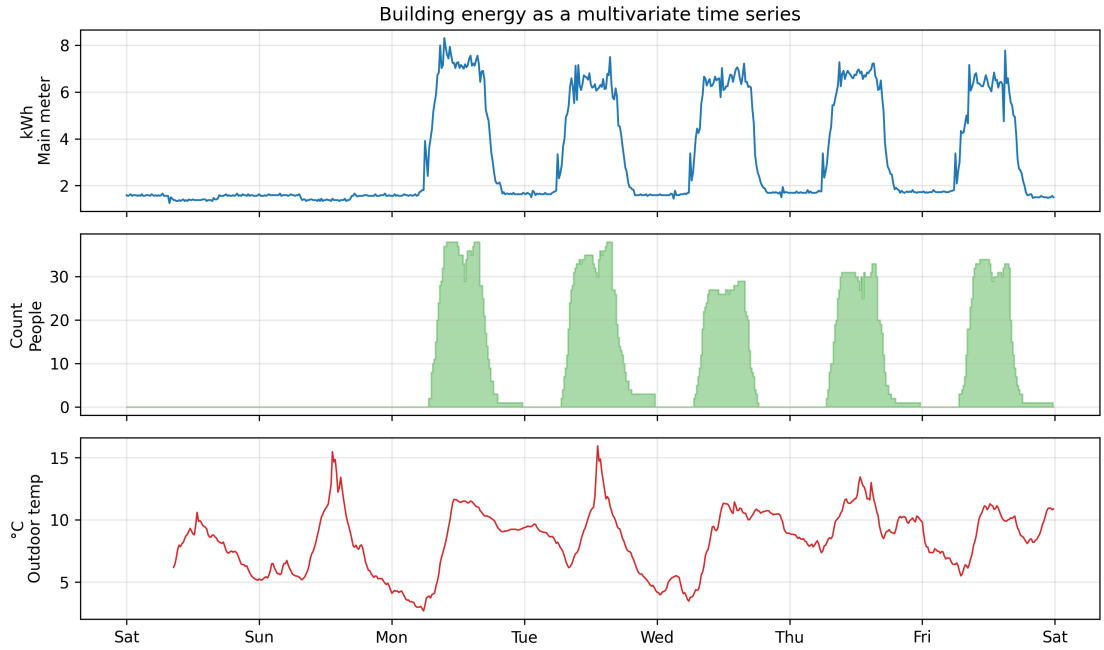


Figure 2.1: Representative multivariate time series showing the main meter load together with occupancy (people count) and outdoor temperature. The plot illustrates how multiple interdependent variables evolve jointly over time.

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 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

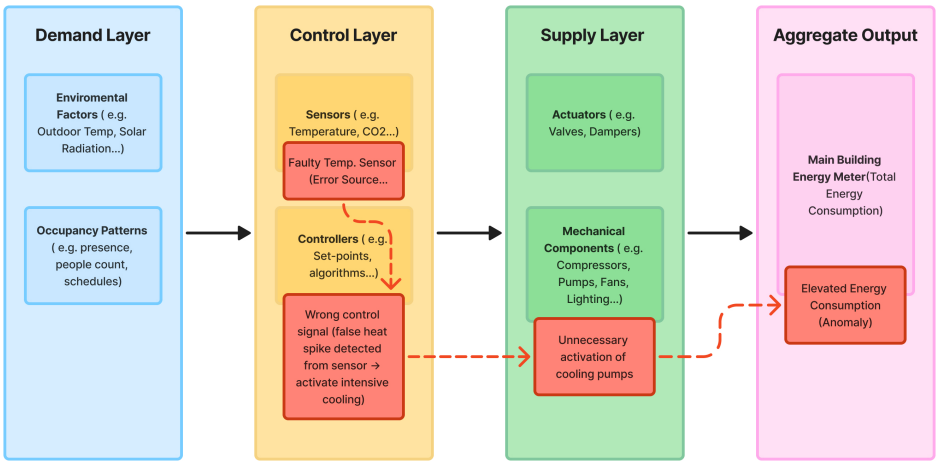


Figure 2.2: Causal chain of building energy consumption from demand over control to supply layer.

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 illustrates this effect on real data: the measured main-meter consumption concentrates in several dense regions at lower loads and exhibits a pronounced right

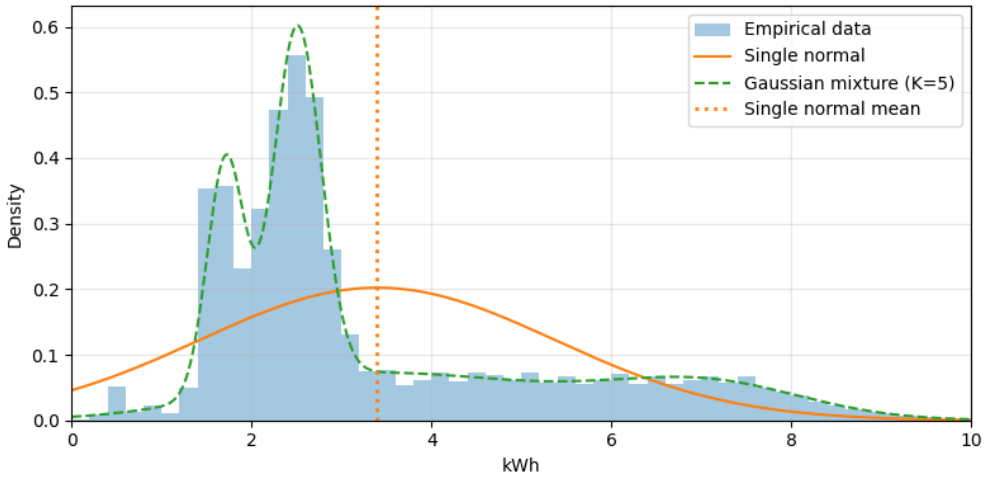


Figure 2.3: Empirical distribution of the building's main meter (15-minute kWh values, histogram) with an overlaid single normal distribution and a Gaussian mixture model with five components, illustrating the mismatch between a unimodal Gaussian model and the multimodal, heavy-tailed structure of real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

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.

2.2. 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.

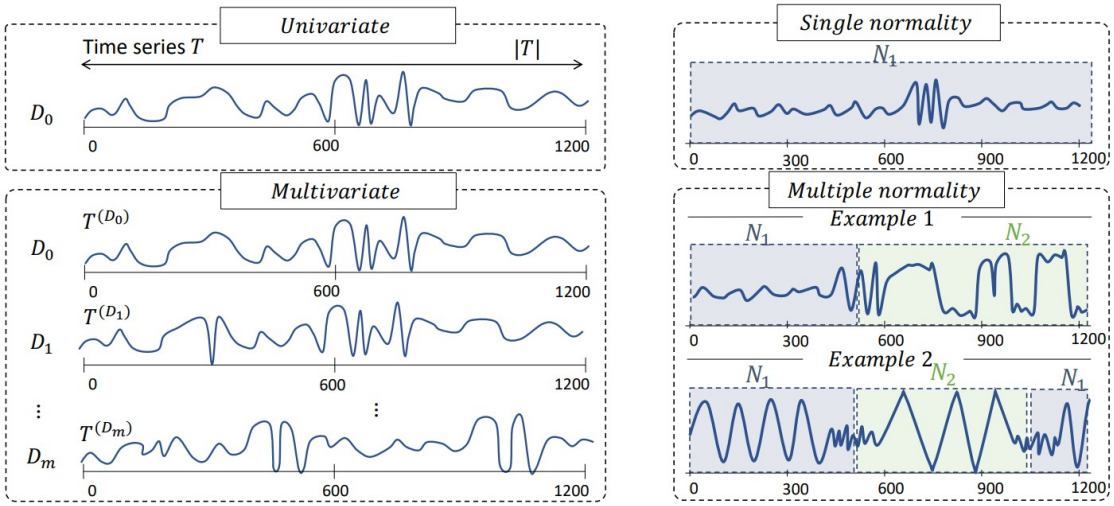


Figure 2.4: Schematic illustration of time series types along two axes—dimensionality (univariate vs. multivariate) and normality regimes (single-mode vs. multi-mode). Adapted from Boniol et al.'s tutorial on new trends in time series anomaly detection [**Boniol2023NewTrends**].

2.2.2. Terminology: Multivariate and Multi-Target Time Series

Throughout this thesis, the term *multivariate* denotes covariate-conditioned time-series modelling. That is, anomaly detection is performed on a single primary energy meter while explicitly conditioning on multiple exogenous driver variables such as weather, occupancy and calendar information. This formulation reflects the standard analytical view in building-energy modelling, where contextual variables are required to resolve causal ambiguity and to distinguish contextual anomalies from physically normal load variations.

In contrast, some anomaly-detection literature uses the term multivariate to describe joint modelling of multiple sensor channels as simultaneous prediction targets. In order to avoid ambiguity, this thesis refers to such settings as *multi-target* (or multi-sensor) time-series modelling.

Accordingly, all experimental investigations in this work address single-target, covariate-conditioned contextual anomaly detection rather than joint multi-target anomaly detection.

2.2.3. Structural Classes of Anomalies

Point Anomalies: Individual observations that deviate from expected behaviour.

Global Point Anomalies: Deviations outside the global historical range.

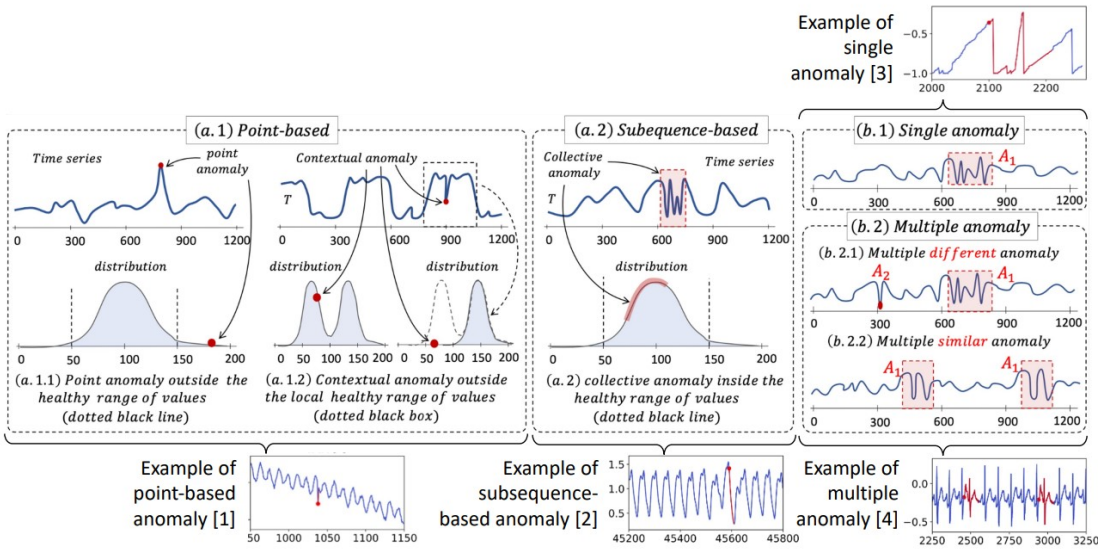


Figure 2.5: Taxonomy of time series anomalies along structural and multiplicity dimensions, distinguishing global and contextual point anomalies, subsequence-based anomalies, and their occurrence as single, multiple different, or multiple similar events. Adapted from Boniol et al.'s tutorial on new trends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

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

Subsequence Anomalies: Deviations manifested through abnormal temporal patterns.

2.2.4. 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.

2.3. 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.

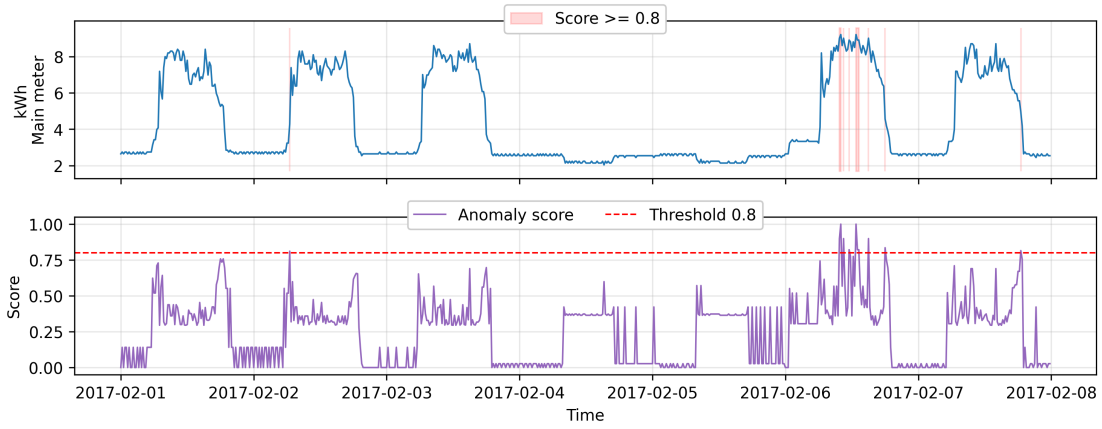


Figure 2.6: Example of an anomaly score $s_i \in [0, 1]$ aligned with the underlying time series. A threshold of 0.8 separates normal points from those flagged as anomalous.

2.3.2. Learning Paradigms

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

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:

Distance-Based: Identify anomalous subsequences by pattern dissimilarity.

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

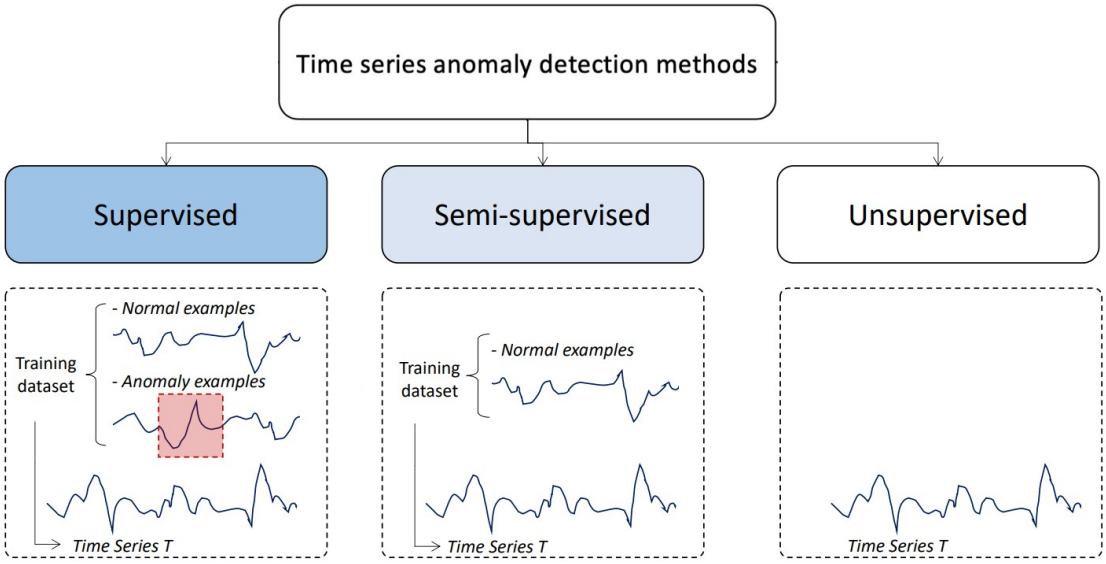


Figure 2.7: Schematic overview of supervised, semi-supervised, and unsupervised learning paradigms for anomaly detection, adapted from Boniol et al.'s tutorial on new trends in time series anomaly detection [**Boniol2023NewTrends**].

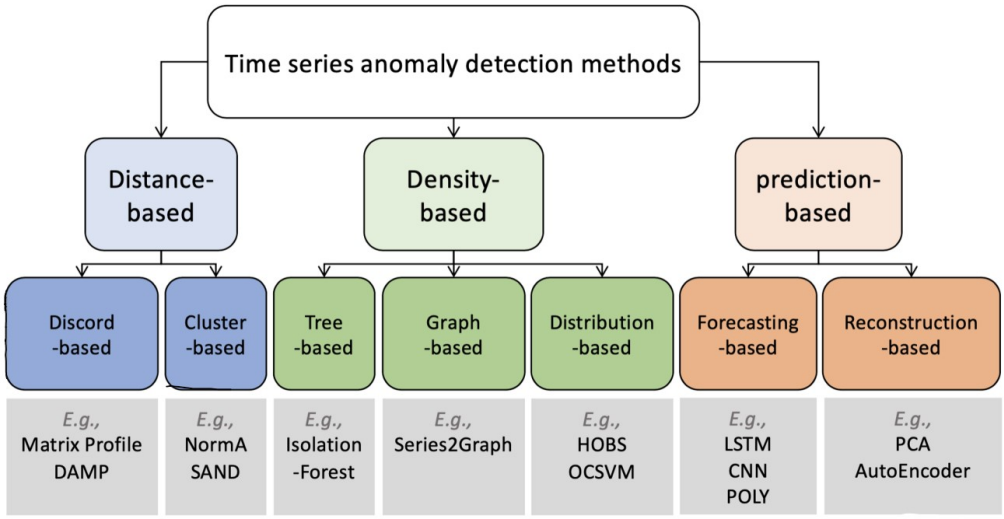


Figure 2.8: Hierarchical taxonomy of anomaly detection methods grouped into distance-based, density-based, and prediction-based families, adapted from Boniol et al.'s tutorial on new trends in time series anomaly detection [**Boniol2023NewTrends**].

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

Benchmarking evaluates anomaly detection performance against labelled reference datasets by comparing model decisions to ground-truth annotations.

2.4.1. Binary Labels and Confusion Matrix

Each observation is assigned a binary label (normal vs. anomalous). Model predictions are evaluated using the confusion matrix, yielding counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

2.4.2. Evaluation Metrics

Performance is summarized using precision, recall, and the F1 score:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

These metrics quantify alarm reliability, detection sensitivity, and their balanced trade-off, respectively.

2.5. Synthesis of Foundations

Building-energy telemetry constitutes a multivariate, multimodal, and non-stationary stochastic process governed by a causal chain spanning demand, control, and me-

chanical execution. Deviations observed at aggregate meters therefore represent manifestations of upstream technical or behavioural faults rather than isolated numerical outliers.

From the structural and statistical properties of these data, the following formal requirements for anomaly detection follow:

1. **Probabilistic multivariate modeling**
   Expected behaviour must be represented as a multivariate mixture distribution rather than a deterministic point estimate, in order to capture multimodal operating regimes and cross-variable dependencies.

2. **Robustness to temporal dependence and persistence**
   Detection must remain stable under strong autocorrelation, seasonal regime shifts, and long-duration persistence, such that sustained deviations are not absorbed into the learned baseline.

3. **Separation of physical faults and digital artifacts**
   Transmission gaps, aggregation spikes, and other digital artifacts must be distinguished from physical anomalies through explicit preprocessing and data-quality handling.

Furthermore, building-energy telemetry is formally classified as a multivariate, multimode time series dominated by contextual point anomalies. Subsequence faults therefore manifest as contextual point deviations at the aggregation horizons relevant for energy monitoring. Persistent deviations present in historical baselines may be absorbed into learned normality (normality drift), representing a fundamental constraint for semi-supervised and unsupervised detection paradigms.

3

Related Work

3.1. Classical Energy Baseline and Rule-Based Detection

Early work on energy anomaly detection in buildings is dominated by deterministic baselines and expert-driven rule systems that encode normative consumption behaviour explicitly. These approaches originate from energy engineering practice and are widely deployed in building management systems due to their transparency and low computational complexity.

Peña et al. [Peñ+16] present a representative rule-based framework for smart buildings in which energy efficiency indicators are derived from HVAC operation and expert knowledge is formalized into a set of anomaly detection rules using data mining techniques. Their system detects predefined inefficiency patterns based on threshold violations and logical conditions applied to multiple sensor streams. While such approaches provide interpretable diagnostics and are well suited for known fault patterns, they rely on static expert rules and lack adaptability to evolving building behaviour, seasonal regime changes, and unseen anomaly types.

Regression-based baselining methods constitute another classical detection paradigm. Liu and Nielsen [LN16] propose an online regression framework for smart-meter anomaly detection in which expected consumption is estimated via supervised learning models and anomalies are detected as residual deviations. These methods enable scalable real-time detection and support streaming deployment; however, they assume relatively stationary consumption baselines and primarily operate on deterministic point forecasts, limiting their robustness under multimodal and non-Gaussian energy distributions.

Overall, classical rule-based and regression-based approaches establish important foundations for energy anomaly detection, but their deterministic formulation and reliance on static baselines restrict their ability to resolve contextual, regime-dependent, and stochastic deviations that characterize modern building-energy telemetry.

3.2. 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 time series anomaly detection (TSAD) is often an artifact of flawed evaluation practices rather than algorithmic superiority.

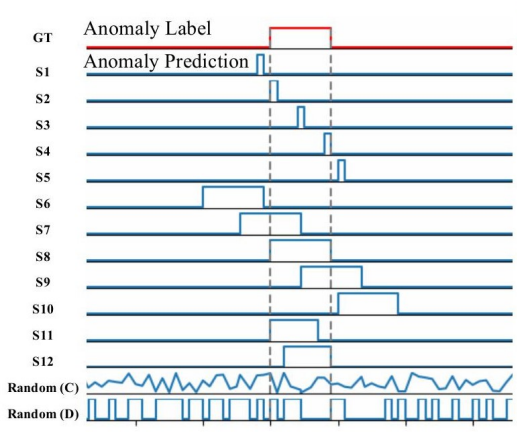
3.2.1. Systemic Flaws and Metric Reliability

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.

3.2.2. Benchmark Evaluation and Model Hierarchy

Evaluation across 1 070 curated time series reveals that statistical methods like subspace principal component analysis (Sub-PCA) [LGW04] 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)) [Wu17] and generative models like stochastic recurrent neural network model OmniAnomaly (OmniAnomaly) [Su+19] consistently outperform statistical baselines in capturing non-



Bias | Indiscrimination | Lack of adaptability

Scenario	Threshold-independent					Threshold-dependent				
	AUC-PR	AUC-ROC	VUS-PR	VUS-ROC	PATE	Standard-F1	PA-F1	Event-based-F1	R-based-F1	Affiliation-F
S1	0.04	0.50	0.13	0.54	0.03	0.00	0.00	0.00	0.00	0.95
S2	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.98
S3	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.99
S4	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.98
S5	0.04	0.50	0.13	0.54	0.18	0.00	0.00	0.00	0.00	0.95
S6	0.04	0.48	0.18	0.62	0.03	0.00	0.00	0.00	0.00	0.93
S7	0.27	0.74	0.61	0.85	0.68	0.50	0.80	0.67	0.55	0.98
S8	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
S9	0.27	0.74	0.61	0.85	0.60	0.50	0.80	0.67	0.55	0.98
S10	0.04	0.48	0.18	0.62	0.14	0.00	0.00	0.00	0.00	0.93
S11	0.81	0.90	0.81	0.90	0.96	0.89	1.00	1.00	0.91	1.00
S12	0.81	0.90	0.81	0.90	0.91	0.89	1.00	1.00	0.91	1.00
Random (C)	0.06	0.56	0.09	0.72	0.06	0.12	0.73	0.21	0.16	0.70
Random (D)	0.04	0.51	0.08	0.66	0.34	0.08	0.15	0.08	0.09	0.68

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].

linear dependencies across multiple sensor channels [Su+19].

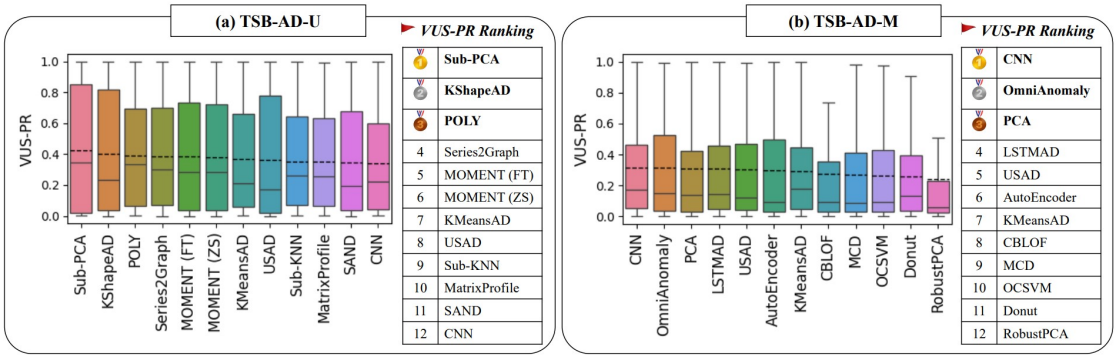


Figure 3.2: Accuracy evaluation of the top 12 methods on (a) univariate (TSB-AD-U) and (b) multivariate (TSB-AD-M) datasets based on the VUS-PR metric. Adapted from Liu and Paparrizos [LP24].

3.2.3. 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. It is important to note that the term *multivariate* in Liu and Paparrizos [LP24] refers to joint multi-sensor anomaly detection, whereas in this thesis it denotes covariate-conditioned detection on a single primary meter. Consequently, the benchmark does not cover multivariate contextual anomaly detection in the sense addressed in this work.

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 adopts the robust evaluation principles and metric recommendations of Liu and Paparrizos [LP24], the experimental design is explicitly adapted to the requirements of multivariate contextual anomaly detection in longitudinal building-energy telemetry, enabling continuous baseline learning under non-stationarity.

3.2.4. Large-Scale Supervised Energy Benchmarks: LEAD 1.0

A prominent large-scale benchmark for energy anomaly detection is LEAD 1.0 [GA22], which provides manually annotated hourly electricity consumption data from 1 413 commercial buildings. Anomalies are labeled based on visually observable deviations from daily and weekly load patterns and include global point anomalies and collective anomalies.

The availability of explicit anomaly labels has enabled supervised classification approaches to achieve extremely high reported detection scores. Recent competition results demonstrate that gradient-boosted tree ensembles combined with extensive change-of-value feature engineering can achieve ROC-AUC values above 0.98 by directly learning the human labeling patterns [Fu22].

However, LEAD 1.0 primarily captures globally visible pattern breaks and does not encode contextual inefficiencies, multivariate causal dependencies, long-term baseline drift, or economic impact semantics. Consequently, supervised models trained on LEAD effectively learn to imitate human visual judgments rather than to detect physically or economically relevant inefficiencies. These properties limit the transferability of LEAD-based detection results to real-world building energy management.

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

The landscape of time-series anomaly detection in energy systems has evolved from classical statistical heuristics toward complex deep learning architectures and, more recently, time-series foundation models. While Morshedi and Matinkhah [MM25] provide a comprehensive survey of convolutional, recurrent, and adversarial neural architectures in IoT anomaly detection, the specific structural properties of building-energy telemetry impose substantially different modeling requirements.

3.3.1. Deep Generative Models and the Advantage of Reconstruction

A central methodological distinction in building-energy research lies between deterministic forecasting models and probabilistic generative models. Azzalini et al. [Azz+25] demonstrate that recurrent autoencoder architectures consistently outperform convolutional variants due to their ability to capture long-range temporal dependencies in sequential meter data.

Within variational autoencoder frameworks, reconstruction probability (RP) has been shown to outperform simple reconstruction error (RE) by explicitly accounting for reconstruction variance, thereby increasing robustness against stochastic fluctuations inherent to building operations. This modeling principle underlies generative architectures such as OmniAnomaly, which employ stochastic recurrent neural networks to learn latent representations of multivariate building telemetry and to characterize normal operational behavior probabilistically [Su+19].

3.3.2. Time-Series Foundation Models in the Energy Domain

Time-series foundation models introduce a paradigm shift by enabling zero-shot and few-shot generalization across heterogeneous datasets. Hela, Handigol, and Arjunan [HHA25] evaluate foundation models such as TimeGPT [GCM23] and MOMENT [Gos+24] on the LEAD 1.0 benchmark, showing that these models exhibit strong zero-shot capabilities for detecting globally visible point anomalies in building-energy time series.

However, benchmark results further indicate that compact generative architectures may outperform foundation models in forecasting-residual-based anomaly detection tasks. Specifically, variational autoencoders trained from scratch surpass large foundation models such as MOMENT [Gos+24] on LEAD 1.0, while Liu and Paparrizos

[LP24] similarly report dominance of lightweight statistical and neural architectures in benchmark-driven TSAD competitions.

Crucially, these findings are confined to deterministic forecasting-residual paradigms and snapshot-based benchmark formulations. Existing benchmarks primarily encode globally visible or subsequence-based anomalies and evaluate models under unimodal Gaussian residual assumptions. They do not represent multivariate contextual inefficiencies, regime-dependent multimodality, long-term baseline drift, or probabilistic deviation semantics that characterize real-world building energy telemetry.

This work therefore departs from the conventional residual-based anomaly detection formulation by treating building-energy anomaly detection as probabilistic deviation from a contextual multivariate baseline. In this regime, foundation models capable of native distributional forecasting—such as Chronos-2—provide architectural capabilities that enable explicit modeling of regime-dependent mixture densities, which are structurally required to represent the multimodal operational states of buildings.

3.3.3. Synthesis of Related Work

The review of established methodologies demonstrates a technical transition from deterministic expert systems toward probabilistic deep learning architectures. Classical rule-based frameworks and regression-based baselining provide high interpretability and low computational complexity but exhibit restricted adaptability to the non-stationary and multimodal characteristics of building telemetry. The assessment of current methodologies is frequently compromised by systemic flaws in existing benchmarks, including mislabeling and biased evaluation metrics such as PA-F1. To resolve these deficiencies, recent research emphasizes robust evaluation standards like VUS-PR and the utilization of deep generative models such as OmniAnomaly. While large-scale supervised datasets like LEAD 1.0 enable high detection scores, they primarily reflect human visual judgments rather than contextual inefficiencies or multivariate causal dependencies. Foundation models such as Chronos-2 offer zero-shot generalization and the capacity for distributional forecasting, yet their application to multivariate contextual detection remains a significant research gap. Consequently, this work departs from conventional deterministic residuals in favor of a probabilistic formulation that explicitly models regime-dependent mixture densities to bridge the identified gap between technical detection and operational remediation.

4

Methodology

This chapter formalizes energy anomaly detection as probabilistic deviation from regime-conditioned multimodal baselines under non-stationary dynamics. It critiques sequential forecasting–based anomaly detection, analyzes the limitations of point and Gaussian prediction under multimodality, and derives distribution-aware scoring mechanisms for mixture density models, together with conservative financial impact quantification and hierarchical localization of anomalous behavior in large-scale building portfolios.

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 implemented as a multi-tenant system supporting simultaneous operation of independent building portfolios within a shared infrastructure. All computational services, storage layers, and frontend applications operate in a logically isolated but physically shared architecture.

The system follows a two-layer design consisting of a backend service layer and a frontend application layer. The backend encapsulates all persistence, processing, and integration services and exposes a unified API to the frontend.

Backend Layer The backend acts as the centralized processing and persistence hub. It manages asset registration, authentication, and data ingestion, hosts the Rule Engine for automated processing, and coordinates specialized microservices for forecasting, anomaly detection, and diagnostics[Eli25d; Eli25e]. All time-series and metadata are stored in a shared PostgreSQL instance extended with TimescaleDB hypertables, enforcing strict logical tenant isolation while enabling centralized resource management and horizontal scalability.

Frontend Layer The frontend provides the primary user interface for all tenants. It exposes dashboards, maps, reports, and analytics for monitoring energy behavior and interacting with the results produced by backend services.

4.1.2. Asset Modeling and Hierarchical Ontology

A central component of the platform is its asset ontology, which provides a structured representation of physical and functional relationships between entities within and across buildings[Eli25c; Eli25a]. Assets are instantiated from reusable templates and organized into complementary hierarchies.

Assets and Templates Assets represent any physical or logical entity, including sensors, rooms, equipment, and buildings. Each asset is instantiated from an Asset Template that predefines semantic attributes such as temperature, occupancy, or power demand, ensuring consistent metadata across tenants and sites.

Dual Hierarchies Assets are organized into two orthogonal trees. The *Local Tree* represents physical containment (e.g., Site $\rightarrow$ Building $\rightarrow$ Floor), while the *Functional Tree* captures technical dependencies (e.g., Heating System $\rightarrow$ Pump $\rightarrow$ Flow Sensor). This dual structure enables both spatial and functional attribution over the same telemetry.

Semantic Tagging Assets are additionally annotated with semantic tags that enable cross-building grouping and multivariate query composition. These tags are used by the anomaly detection pipeline to assemble contextual feature sets and hierarchical diagnostic views across tenants.

4.2. Formal Design Objectives and System Requirements

The anomaly detection framework is designed as a probabilistic, large-scale contextual inference system for non-stationary building telemetry. The following objectives formally define the design constraints and optimization goals governing the detection methodology and system architecture.

4.2.1. Functional Objectives

Contextual Fidelity The system must detect deviations relative to regime-conditioned normative behavior rather than absolute residual magnitudes, ensuring invariance to seasonal, occupancy-driven, and weather-induced regime shifts.

Distributional Validity Expected behavior must be represented through calibrated probabilistic bounds that remain valid under multimodal, heavy-tailed, and non-Gaussian operating regimes.

Drift Robustness Detection sensitivity must be preserved under slow baseline drift and abrupt regime transitions without requiring explicit retraining cycles.

Persistence Preservation Sustained inefficiencies must remain detectable and must not be absorbed into the learned normative baseline.

Hierarchical Diagnosability Detected anomalies must be attributable to concrete physical subsystems through hierarchical asset ontologies.

Economic Interpretability Deviations must be translated into conservative, physically defensible financial loss estimates.

4.2.2. Operational Objectives

Scalability The framework must support horizontally scalable, parallel processing of independent building portfolios under multi-tenant operation.

Multi-Tenant Isolation Detection, scoring, and attribution must operate independently across tenants and building portfolios.

Data Integrity Robustness Transmission gaps, aggregation spikes, and recovery artifacts must be explicitly mitigated to prevent non-physical anomaly generation.

External Driver Separation Exogenous environmental drivers (e.g., weather extremes) must be distinguishable from endogenous technical faults.

4.3. Financial Impact Quantification

Let x_t denote the observed energy consumption and $p_t(x)$ the predictive distribution of nominal operation at time t. Using a single global mean as reference is invalid under multimodal regimes, as the mean may lie in low-density regions that are physically unrealizable.

4.3.1. Distribution-Aware Baseline

If a mixture density is available,

$$p_t(x) = \sum_{k=1}^{K} \pi_{t,k} \mathcal{N}(x \mid \mu_{t,k}, \sigma_{t,k}^2),$$

the baseline is chosen as the mean of the mixture component whose density best explains the observed value, i.e., the closest plausible operating regime:

$$k^* = \arg\max_k \pi_{t,k} \mathcal{N}(x_t \mid \mu_{t,k}, \sigma_{t,k}^2), \qquad \tilde{\mu}_t = \mu_{t,k^*}.$$

The signed deviation

$$\Delta x_t = x_t - \tilde{\mu}_t$$

represents instantaneous excess ($\Delta x_t > 0$) or savings ($\Delta x_t < 0$) relative to the regime-consistent baseline.

4.3.2. Fallback Without Mixture Information

If only unimodal uncertainty is available, the baseline is conservatively shifted by one standard deviation in the direction of the observation:

$$\tilde{\mu}_t = \mu_t + \sigma_t.$$

4.3.3. Economic Impact

The monetary impact is defined as

$$\Delta C_t = \Delta x_t \cdot c_t,$$

where c_t denotes the unit energy price. Positive values correspond to avoidable cost, while negative values represent verifiable savings.

This formulation yields conservative, regime-consistent and economically interpretable impact estimates under non-Gaussian and non-stationary operating conditions.

4.4. Hierarchical Root Cause Analysis and Action Synthesis

Aggregate anomaly detection provides limited operational value unless deviations can be localized to concrete physical subsystems. The proposed framework therefore performs hierarchical root cause analysis (RCA) using the ontological asset model of the Eliona platform.

4.4.1. Ontology-Guided Attribution

All building assets are organized in dual hierarchies representing both geographical containment (e.g., site, building, floor, room) and functional decomposition (e.g., main meter, sub-meter, device). When an anomaly is detected at an aggregate level, all subordinate assets are queried for their individual anomaly impact.

Assets are ranked by their cumulative contribution over the anomaly interval. Those with the highest contribution are identified as the most probable physical sources of the aggregate deviation. This enables localization not only to specific meters but to concrete functional subsystems and physical zones.

4.4.2. Aggregation by Asset Type

To detect systematic inefficiencies, anomaly impact is additionally aggregated by semantic asset categories (e.g., HVAC, lighting, plug loads). This reveals distributed but structurally consistent inefficiencies that may not be visible at individual device level.

4.4.3. Contextual Synthesis and Action Generation

Localized assets, aggregated impacts, and contextual conditions (time, calendar, weather) are consolidated into a structured diagnostic representation and passed to a domain-specialized large language model (LLM).

The LLM operates exclusively at the interpretation layer. It converts structured evidence into human-readable explanations and actionable operational recommendations, including quantified savings estimates.

4.4.4. Design Rationale

All detection, scoring, and financial quantification are performed deterministically from measured data and building ontologies. The LLM is used exclusively for semantic interpretation and recommendation synthesis and does not influence anomaly detection or decision boundaries, ensuring traceability, auditability, and regulatory compliance.

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 $(x_{t-w}, \ldots, x_{t-1})$ 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 4.1 illustrates this phenomenon using a model trained with a 24-hour historical window plus time-based features (time of day, `is_weekend`).

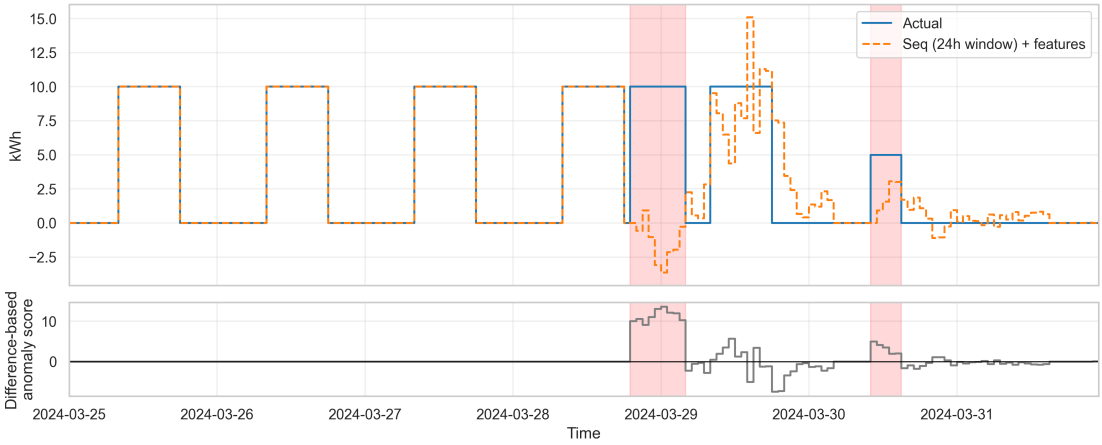


Figure 4.1: Prediction behavior of a model using a 24 h historical window plus time features. The top panel shows actual vs. predicted values; the bottom panel shows the difference-based anomaly score. Note the erratic predictions even after the anomaly ends as the unseen data propagates through the sliding window.

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.

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

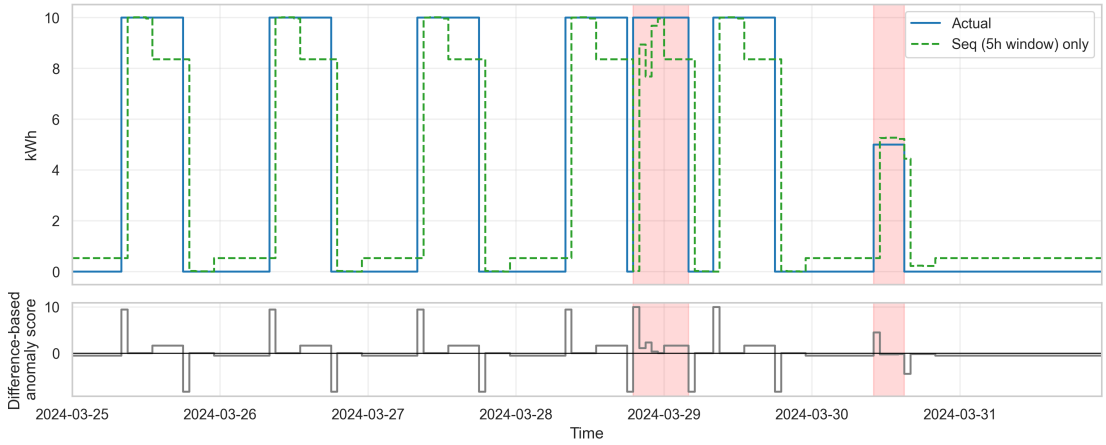


Figure 4.2: Prediction behavior of a model using only a short (5-step) historical window. The model correctly identifies the onset of anomalies but rapidly adapts to the new level, causing the anomaly score to drop back to near zero while the anomaly is still ongoing.

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 4.3 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.

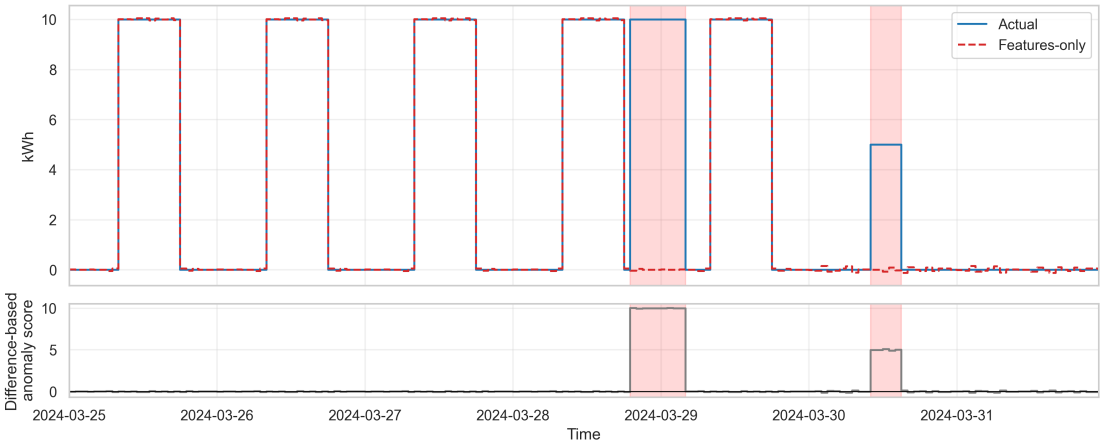


Figure 4.3: Behavior of a features-only model (no historical consumption input). The prediction relies solely on context (time/weekend), resulting in a stable baseline and accurate detection of sustained anomalies without adaptation.

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 4.4 applies this logic to the unstable 24-hour window model from Figure 4.1.

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.

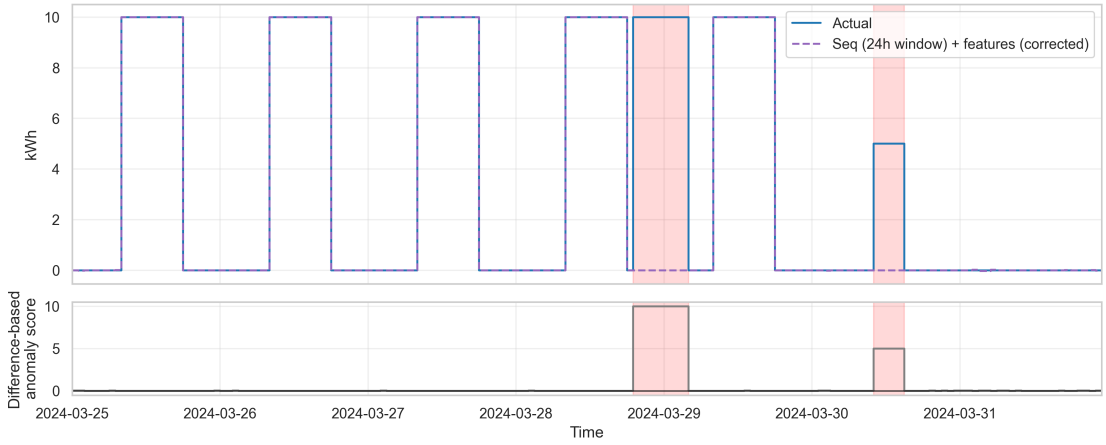


Figure 4.4: The same 24-hour window model from Figure 4.1, but applied with inference-time imputation. When an anomaly is detected, the predicted value replaces the actual value in the sliding window for future steps. This prevents error propagation and maintains a high anomaly score throughout the event.

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 4.5 shows that all three model families—deterministic dense regression, single-Gaussian prediction, and Mixture Density Networks (MDN)—deliver visually similar means, yet their anomaly scoring behavior diverges drastically.

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

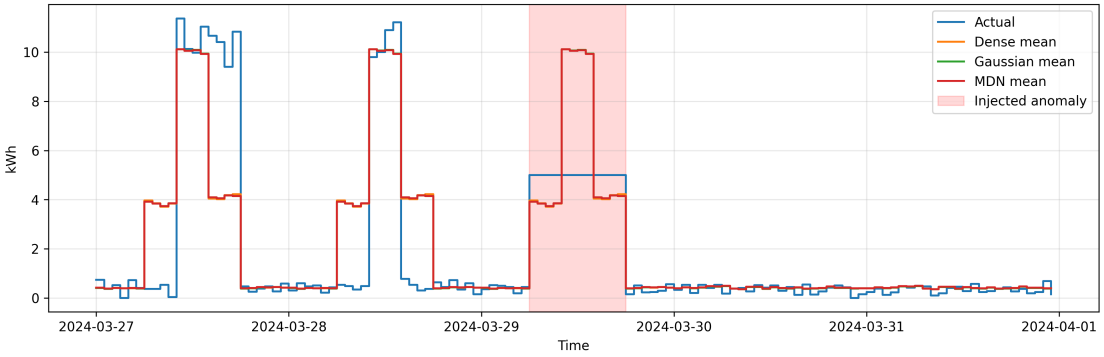


Figure 4.5: Predicted means over the Variable Shift horizon. Dense, Gaussian, and MDN models track the two regimes, masking the scoring deficiencies discussed in Sections 4.6.1–4.6.3.

matches the biased mean. The residual trace in Figure 4.6 exposes this contradiction: the absolute error balloons whenever the device operates normally, yet it contracts when the genuine anomaly occurs.

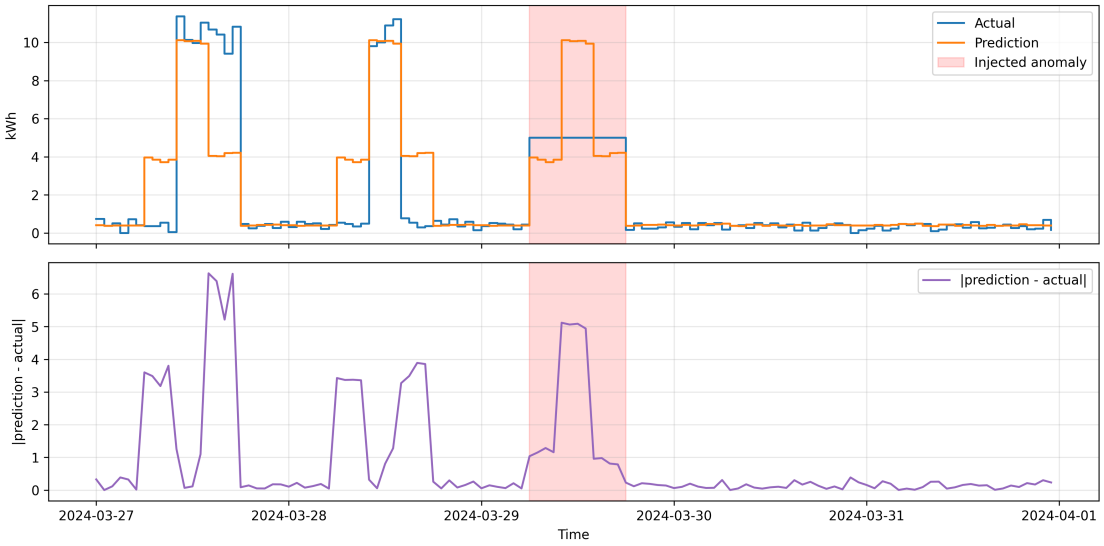


Figure 4.6: Dense regressor residuals over the Variable Shift dataset. The mid-range prediction inflates anomaly scores for legitimate operating states, while the stuck-at-5 fault yields a small residual.

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 4.7) 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 4.8 makes the distortion visible: the green, high-probability band spans the median instead of the true modes.

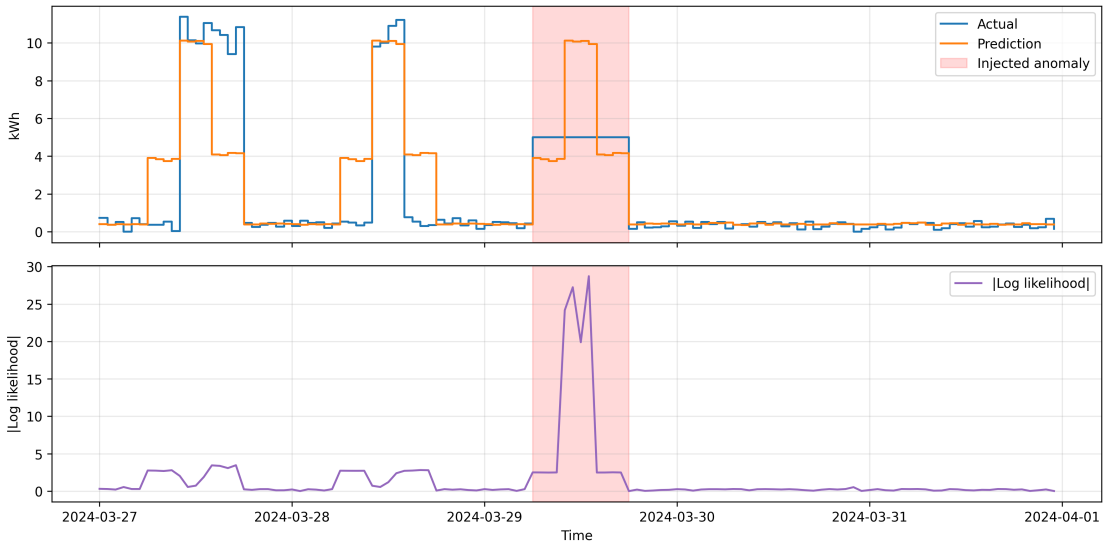


Figure 4.7: Absolute log-likelihood trace for the single-Gaussian predictor. The stuck-at-5 anomaly aligns with the high-likelihood center, suppressing the score.

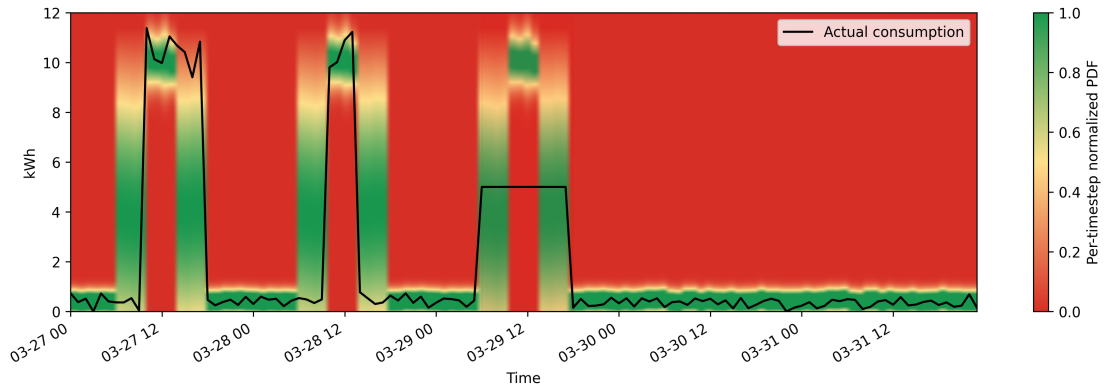


Figure 4.8: Per-timestep normalized PDF for the Gaussian model. High probability mass accumulates between the actual clusters, illustrating the variance-stretching paradox.

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 4.10 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 4.9). 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.

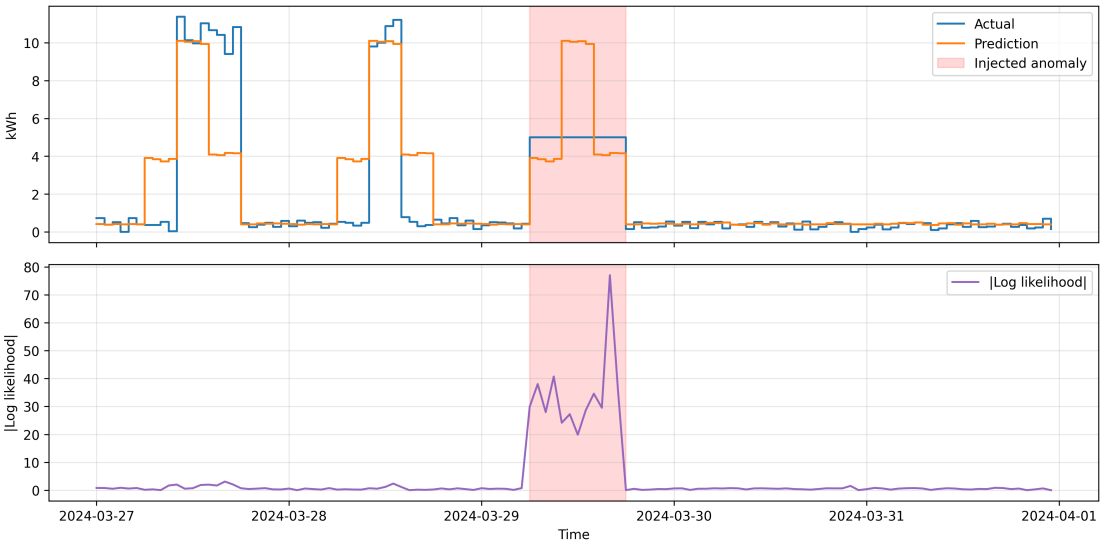


Figure 4.9: MDN absolute log-likelihood trace. The stuck-at-5 anomaly triggers a sustained spike because the value resides in a low-probability region between mixture components.

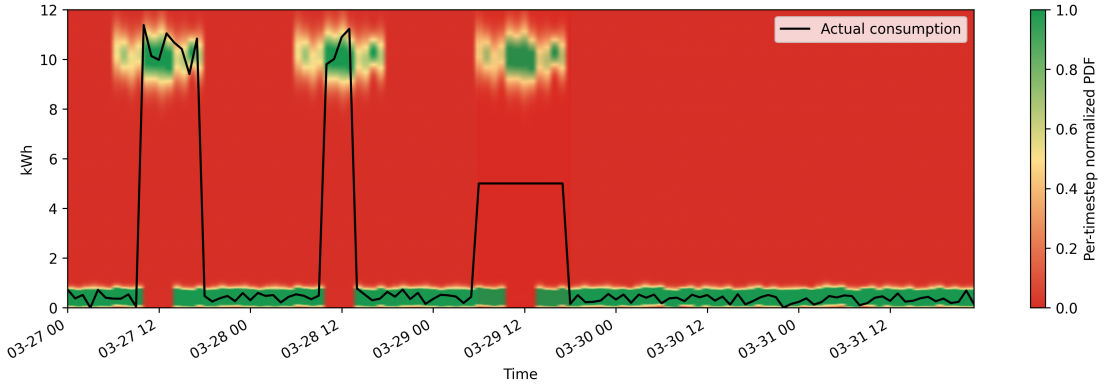


Figure 4.10: MDN normalized PDF heatmap. Two distinct high-probability ridges align with the real operating modes, while the middle band remains improbable.

4.7. Distribution-Aware Anomaly Scoring for Mixture Density Models

Building on Section 4.6, which details the Variable Shift dataset and its bimodal operating regimes, we now analyze how distribution-aware anomaly scores behave when the predictive density itself is multimodal. The deterministic residual failures from Section 4.5 are amplified in this setting because no single "expected value" exists, making deviation from the mean a misleading proxy for abnormality.

Figure 4.11 extends the MDN perspective by overlaying several candidate scores derived from the same mixture distribution: mean residuals, PIT, negative log-likelihood, and the proposed Density–Quantile (DQ) family.

4.7.1. Mean Residual: Failure Under Multimodality

The most common anomaly score is the absolute residual between the observation x_t and the predicted mean μ_t:

$$s_t^{\text{mean}} = |x_t - \mu_t|. \tag{4.1}$$

In multimodal settings, the mean of the predictive distribution often lies in a region of *low probability mass*. As shown in Figure 4.11, the MDN mean converges to an intermediate value between the two legitimate modes. Consequently, observations that are perfectly normal but belong to either mode exhibit large residuals and are falsely flagged as anomalous. Conversely, the injected anomaly—located near the mean but inside a low-density valley—produces a small residual and is incorrectly classified as normal.

This demonstrates that residual magnitude is not a valid proxy for abnormality when the expected behaviour cannot be represented by a single point estimate.

4.7.2. Probability Integral Transform (PIT)

A distribution-aware alternative is the Probability Integral Transform (PIT), which maps each observation to its cumulative probability under the predicted distribution:

$$\text{PIT}_t = F_t(x_t) = \int_{-\infty}^{x_t} p_t(y)\,\mathrm{d}y, \tag{4.2}$$

Figure 4.11: Comparison of anomaly scoring methods derived from a Mixture Density Network under a bimodal operating regime with an injected intermediate anomaly. The top panel shows the predicted probability density together with the actual observation and the MDN mean. Subsequent panels compare mean residuals, PIT-based scores, negative log-likelihood, and the proposed Density–Quantile (DQ) scores and severities.

where $p_t(y)$ denotes the MDN predictive density at time t. A symmetric anomaly score can be defined as:

$$s_t^{\text{PIT}} = 1 - \text{PIT}_t. \tag{4.3}$$

PIT correctly identifies observations in the extreme tails of the distribution. However, it remains insensitive to *low-density regions between modes*. In Figure 4.11, the injected anomaly lies near the median of the distribution and therefore yields a moderate PIT value, despite being highly unlikely. PIT thus fails to detect anomalies that occupy density valleys rather than tails.

4.7.3. Negative Log-Likelihood and Its Limitations

Another principled score is the negative log-likelihood (NLL):

$$s_t^{\text{NLL}} = -\log p_t(x_t). \tag{4.4}$$

NLL correctly assigns high anomaly scores to observations in low-density regions, including the valley between modes. As shown in Figure 4.11, it robustly detects the injected anomaly.

However, NLL values are *not comparable across time*. Each timestamp t corresponds to a different predictive distribution with different entropy, variance, and scale. As a result, absolute NLL magnitudes cannot be meaningfully thresholded or aggregated over time, limiting their use for persistence analysis, severity ranking, and financial quantification.

4.7.4. Density–Quantile (DQ) Probability

To obtain a score that is both distribution-aware and comparable across time, this work introduces the Density–Quantile (DQ) probability. Instead of evaluating the density at a single point, DQ measures the proportion of probability mass that is *less likely* than the observed value:

$$\text{DQ}_t = \int_{\{y : p_t(y) \le p_t(x_t)\}} p_t(y)\, \mathrm{d}y. \tag{4.5}$$

By construction, $\text{DQ}_t \in (0, 1]$ and is invariant to the shape, scale, and entropy of the underlying distribution. Observations in high-density regions yield large DQ values, while points located in tails or low-density valleys yield small values.

An anomaly score can therefore be defined as:

$$s_t^{\text{DQ}} = 1 - \text{DQ}_t. \tag{4.6}$$

As shown in Figure 4.11, this score simultaneously suppresses false positives for legitimate operating modes and sharply highlights the injected anomaly located between the modes.

4.7.5. Density–Quantile Severity Scaling

While $1 - \mathrm{DQ}_t$ provides a normalized anomaly score, it does not reflect the *relative improbability* of extreme events. For example, the difference between $\mathrm{DQ} = 0.99$ and $\mathrm{DQ} = 0.98$ corresponds to a doubling of unlikeliness, yet both values are close on a linear scale.

To address this, a severity transformation is introduced:

$$\mathrm{Severity}_t = \min\left(1,\ \frac{p_{\min}}{\mathrm{DQ}_t}\right), \tag{4.7}$$

where $p_{\min}$ defines the minimum reference quantile that maps to maximum severity.

This transformation preserves the ordering induced by DQ while amplifying differences in the extreme low-probability regime. By selecting $p_{\min}$, the sensitivity of the detector can be explicitly controlled, as illustrated in Figure 4.11 for $p_{\min} = 10^{-2}$ and $p_{\min} = 10^{-4}$.

4.7.6. Summary

Density–Quantile scoring combines the strengths of likelihood-based detection with the comparability of quantile methods. Unlike residuals, it respects multimodality; unlike PIT, it captures density valleys; and unlike NLL, it produces normalized, time-comparable scores suitable for persistence tracking, severity ranking, and downstream financial impact estimation. For these reasons, DQ-based scoring forms the core anomaly quantification mechanism in this work.

4.8. Methodological Scope

The proposed methodology formalizes energy anomaly detection as probabilistic deviation from regime-conditioned multimodal baselines under non-stationary dynamics and maps these deviations to economically interpretable and diagnostically localizable events.

5

Benchmarking

5.1. Benchmark Design and Dataset Generation

To evaluate detection robustness under realistic non-stationary building dynamics, a synthetic benchmark dataset was generated with the Building Optimization Performance Test Framework (BOPTEST)[**BOPTEST**]. The benchmark isolates physical building behavior from telemetry artifacts, providing a controlled yet realistic evaluation environment.

A full year of 15-minute resolution data was produced from the *Multizone Office Complex Air* test case, which represents a large office building with coupled thermal zones, air-handling units, and realistic occupancy schedules. This annual baseline captures seasonal dynamics and supports out-of-season generalization tests while avoiding sensor dropouts or undocumented control overrides.

5.2. Feature and Target Definition

The exported telemetry was decomposed into contextual drivers and consumption targets. Control-loop variables were explicitly excluded to prevent fault leakage into the feature space, ensuring that anomalies remain detectable through residual behavior rather than direct control-state exposure.

Contextual features include weather variables, occupancy counts, calendar indicators, and cyclical encodings of diurnal and annual patterns. Seventeen electrical meters

were treated as targets, spanning the aggregate main meter alongside critical HVAC, pumping, and lighting subsystems to support both financial attribution and hierarchical diagnostics.

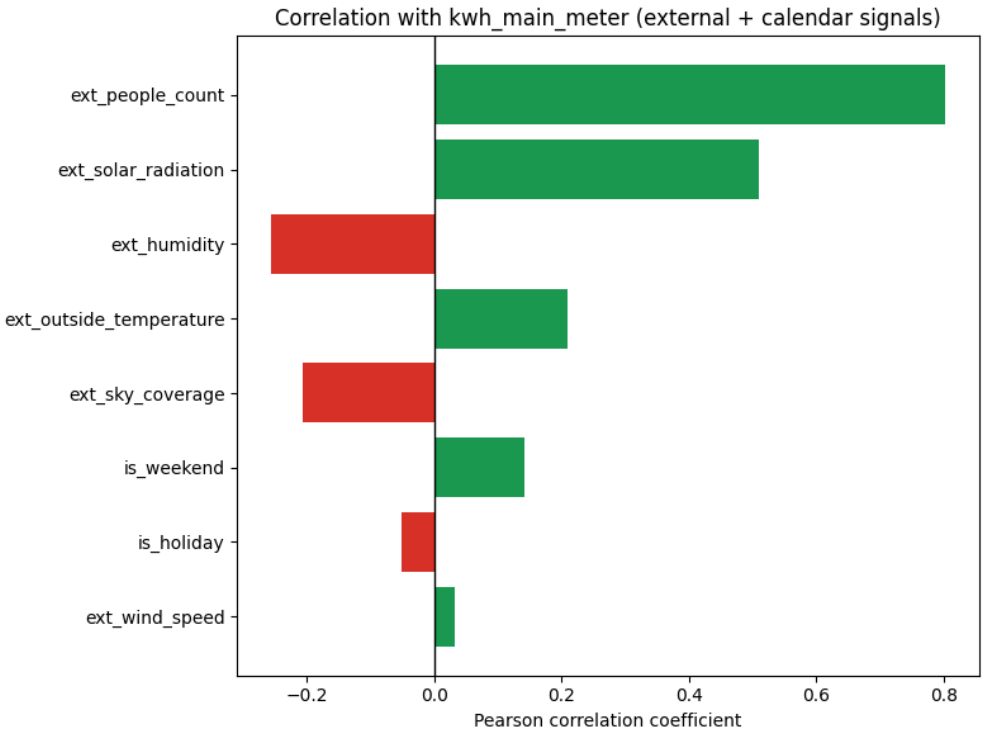


Figure 5.1: Empirical Pearson correlation of contextual and external drivers with the main electricity meter over the synthetic 2007 benchmark year.

5.3. Data Segmentation and Anomaly Injection

The annual baseline was segmented into multiple training–evaluation regimes to assess sample efficiency and seasonal generalization, including six-month, three-month, and two-week windows. Injected anomalies were constrained to remain below 5% of total points to preserve realistic anomaly prevalence on the main meter and sub-meters.

Synthetic perturbations emulate realistic operational, control, and contextual fault classes such as sustained drifts, extended deviations, off-hours activity, pattern shifts, and localized spikes. Each anomaly is labeled with onset, duration, and magnitude to support fine-grained evaluation of detection latency, persistence coverage, and financial attribution.

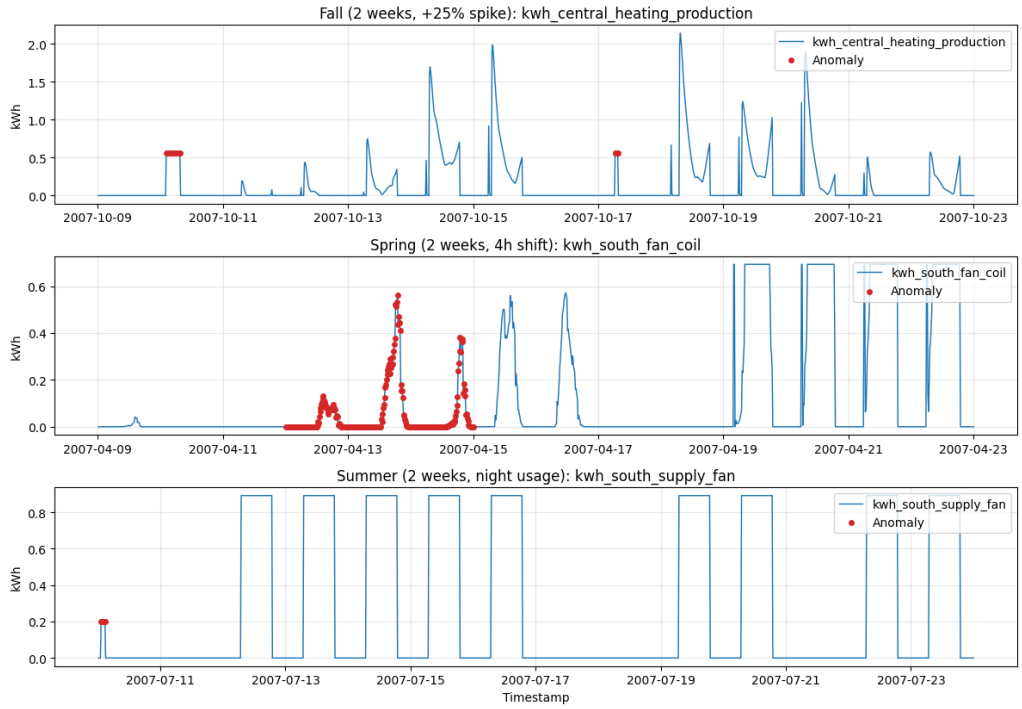


Figure 5.2: Representative sub-meter excerpts from the curated benchmark slices (fall spike, spring pattern shift, summer off-hours). Red markers indicate anomaly windows where the targeted device deviates from its baseline regime.

5.4. Evaluation Constraints and Benchmark Limitations

While the BOPTEST-based benchmark enables controlled and reproducible evaluation, several constraints affect the interpretation and comparability of the resulting metrics.

5.4.1. Training Stability and Coverage Bias

Several evaluated architectures require extensive per-meter hyperparameter tuning. Mixture-density-based models in particular exhibit high sensitivity to initialization and regularization, leading to unstable convergence on subsets of meters.

As a consequence, benchmark coverage differs substantially between methods. Aggregate scores for unstable models are computed over reduced subsets of meters, introducing implicit coverage bias. Reported performance therefore reflects both detection accuracy and training robustness under limited tuning budgets.

5.4.2. Comparability Across Model Classes

Trainable models are evaluated on fixed-length baseline windows augmented with synthetic anomalies. Foundation models, in contrast, condition on historical context preceding the evaluation window, resulting in unequal informational priors at inference time. This discrepancy is most pronounced in short-window and seasonal transfer experiments.

The season-matched three-month baseline configuration provides the closest approximation of comparable operating conditions. In this configuration, Chronos-2 can fully condition on the complete baseline history within its maximum context length, eliminating structural disadvantages relative to trainable models.

5.4.3. Interpretation Scope

Benchmark metrics should therefore be interpreted as indicators of practical deployability and robustness under realistic engineering constraints rather than as absolute measures of algorithmic superiority. Emphasis is placed on persistence tracking, qualitative behavior, and economic interpretability rather than raw aggregate scores alone.

5.5. Comparative Model Performance and Structural Evaluation

A total of up to 16,979 benchmark runs per model were executed across all dataset slices. Classical CNN baselines adapted from the TSB-AD benchmark perform poorly on building-energy telemetry despite strong multivariate results in generic TSAD benchmarks, confirming that autoregressive input windows destabilize anomaly detection under non-stationary operating regimes.

A CNN Feature-Only architecture that relies solely on exogenous contextual drivers achieves markedly higher detection scores, indicating that contextual baselines provide superior normative representations for building-scale telemetry.

5.5.1. Stochastic and Hybrid Architectures

Pure mixture density networks underperform deterministic contextual models. In contrast, the hybrid `CNN_MDN` delivers the highest VUS-PR on long-horizon baselines, while residual dense networks dominate short-horizon regimes. Detection accuracy increases

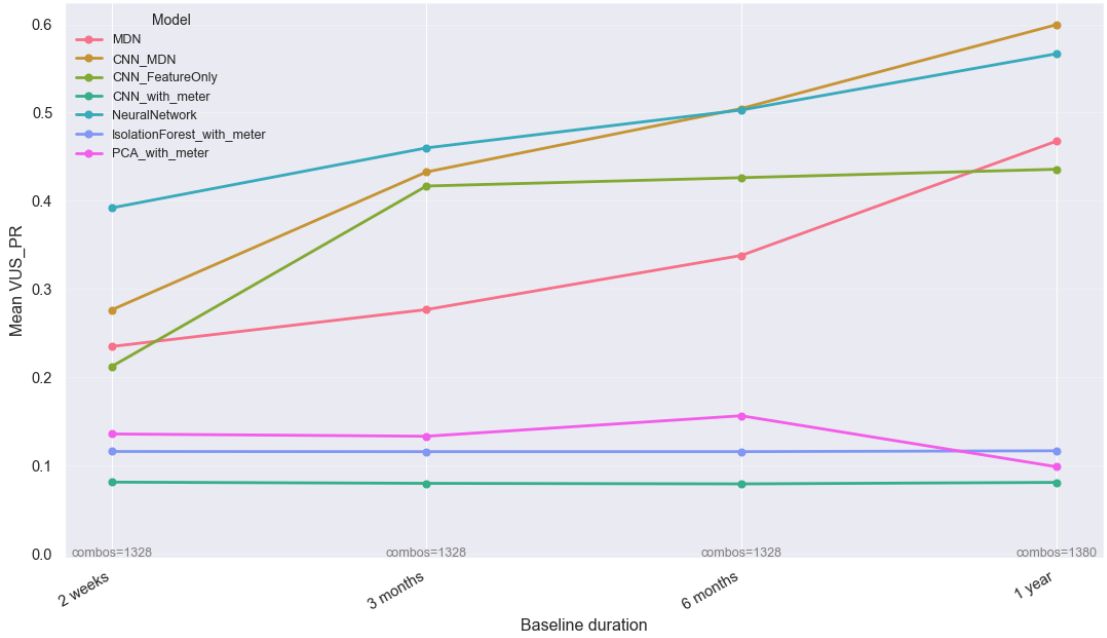


Figure 5.3: Comparison of neural baselines and mixture-density variants across benchmark slices.

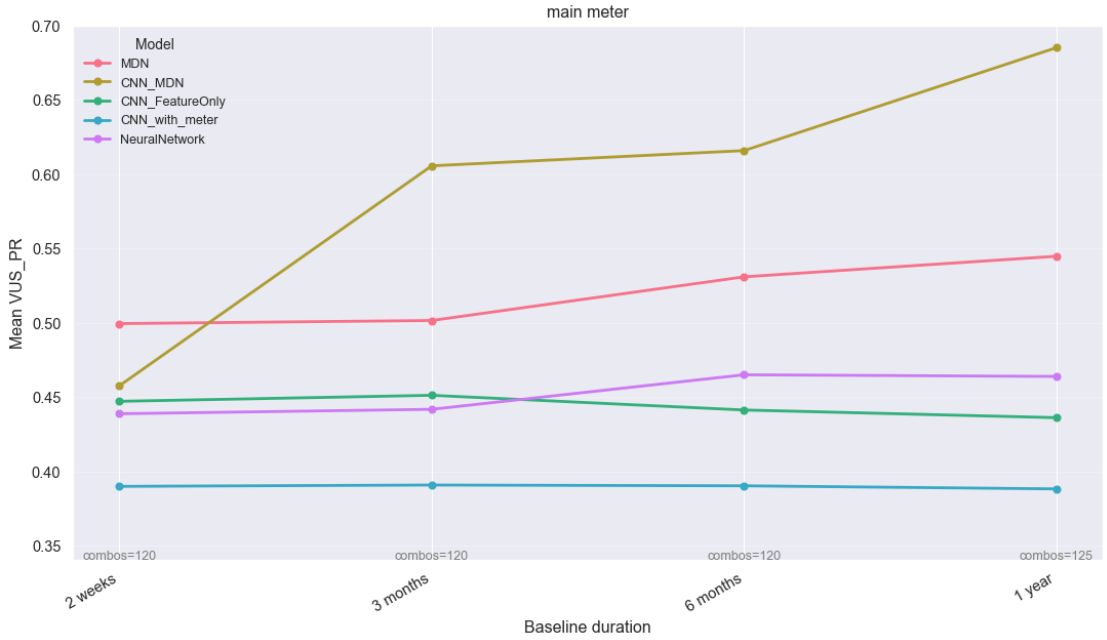


Figure 5.4: Benchmark comparison on the building main meter, highlighting feature-only and stochastic output effects.

monotonically with the amount of baseline context available for training.

5.5.2. Training Stability and Classical Baselines

MDN layers exhibit sensitivity to initialization and regularization, leading to coverage gaps across meters. PCA and Isolation Forest baselines consistently yield the lowest performance, confirming that linear and tree-based statistical models fail to capture the non-linear contextual dependencies present in building-energy telemetry. The consolidated training history remains available in Appendix A.1 (Figure A.1).

5.5.3. Season-Matched Three-Month Evaluation

To minimize seasonal confounding, models were compared under a season-matched three-month baseline. While `CNN_MDN` attains the strongest mean VUS-PR, deterministic neural baselines remain competitive on several anomaly classes. Chronos-2 trails the strongest trainable models, with fine-tuning improving primarily off-hours detection. PatchTST and OmniAnomaly underperform, whereas Autoformer shows strong performance on spike-like anomalies.

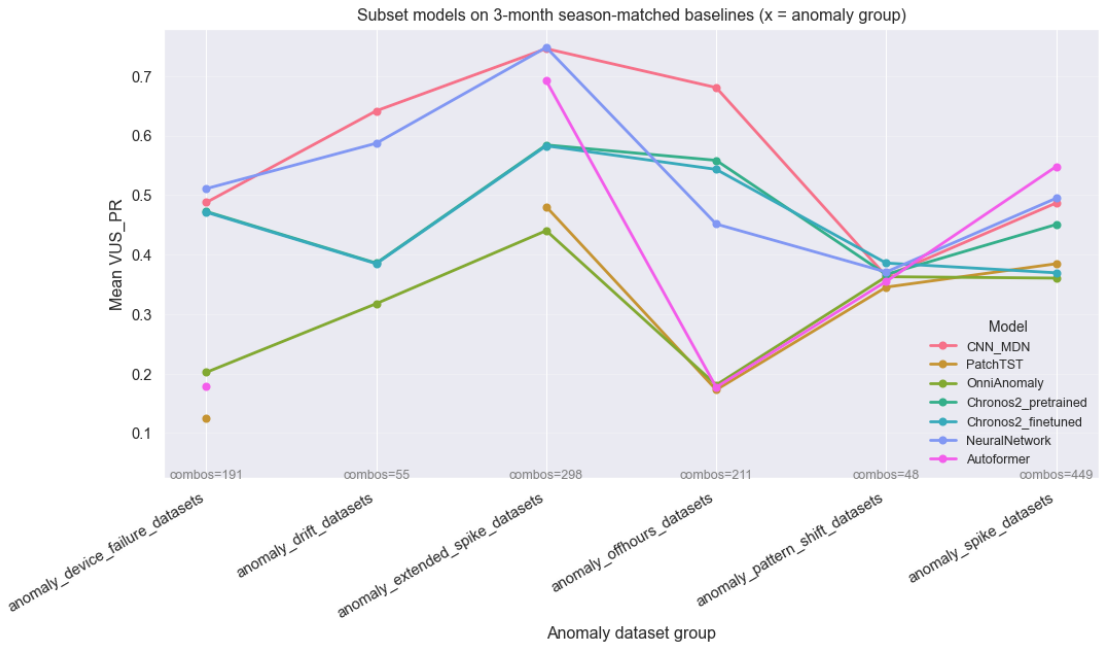


Figure 5.5: Chronos-2 vs. `CNN_MDN` under season-matched three-month baselines.

5.5.4. Seasonal Translation Sensitivity

Mean VUS-PR degrades monotonically as the seasonal distance between training and evaluation windows increases. The largest performance losses are observed for `CNN_MDN` and PatchTST, while Autoformer remains comparatively stable across seasonal shifts.

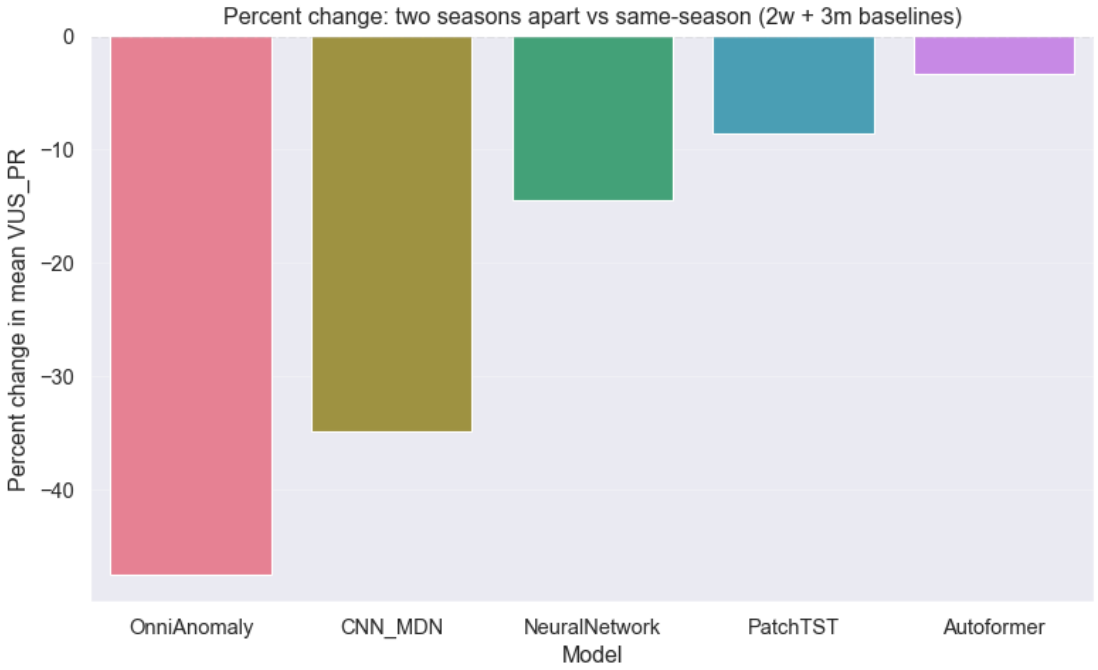


Figure 5.6: Seasonal translation sensitivity of mean VUS-PR versus seasonal distance.

5.5.5. Model Selection Rationale

Although `CNN_MDN` attains the highest mean VUS-PR in most benchmark configurations, Chronos-2 is selected as the primary deployment model due to superior operational scalability and robustness.

Trainable baselines require explicit per-meter fitting and periodic retraining to track distributional drift, implying millions of managed models and fragile retraining pipelines in large building portfolios. In contrast, Chronos-2 operates in a zero-shot regime and adapts online by conditioning on recent context, enabling immediate assimilation of structural regime changes (e.g., HVAC retrofits or occupancy shifts) without retraining orchestration or warm-up phases.

Chronos-2 therefore provides a maintenance-minimal and drift-robust solution that remains competitive in detection accuracy while being operationally superior in portfolio-

scale deployments.

While Chronos-2 does not model explicit multimodal densities, it produces calibrated distribution-free quantile bounds. This enables PIT-based anomaly scoring under non-Gaussian and heavy-tailed regimes while avoiding the instability and overfitting tendencies observed in mixture-density architectures.

6

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.

6.1. 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.

6.1.1. 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.

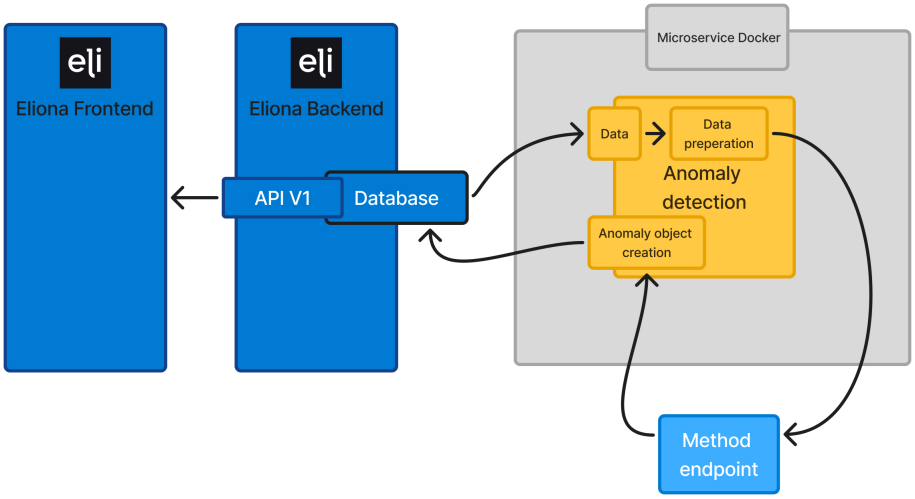


Figure 6.1: High-level deployment of the anomaly detection microservice and Python analytics endpoint within the Eliona and Azure ecosystems.

6.1.2. 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.

6.2. Chronos-2 Analytics Endpoint

The analytical core is implemented as a dedicated Python microservice hosting the Chronos-2 transformer model. The service processes batched multivariate time se-

ries, receiving historical context and exogenous covariates and returning probabilistic quantile forecasts that define the regime-conditioned normative envelope.

For production deployment, the service is exposed as a managed online endpoint in Azure Machine Learning. This provides automated horizontal scaling, versioned model management, and containerized deployment, enabling seamless operation in both cloud and on-premise environments while supporting non-stationary adaptation requirements.

6.3. Scala Microservice: Multi-Tenant Orchestration

System-wide orchestration is implemented as a dedicated Scala microservice that coordinates data ingestion, contextual enrichment, anomaly scoring, and tenant-level isolation. Predictive inference is delegated to the Python Chronos-2 endpoint, while the Scala layer performs all asynchronous I/O, scheduling, and hierarchical orchestration.

Scala is selected for its strong static typing and high-throughput concurrency model, enabling safe parallel processing of large multivariate telemetry tensors and complex building ontologies under multi-tenant operation.

6.3.1. Multi-Tenant Lifecycle Management

Tenants are managed through a centralized orchestration layer that enforces strict logical isolation across organizations. Independent worker pipelines are instantiated per tenant and operate on disjoint telemetry scopes, ensuring that ingestion, processing, and anomaly attribution remain isolated across building portfolios.

Tenant lifecycles are synchronized periodically with licensing and configuration state. Worker pipelines are automatically instantiated, restarted, or terminated in response to tenant state changes, ensuring fault tolerance and consistent multi-tenant operation without manual intervention.

6.4. Data Acquisition and Processing Pipeline

Raw building telemetry is transformed into structured model inputs through a multi-stage pipeline that performs attribute selection, contextual enrichment, and robust data conditioning.

6.4.1. Attribute Selection and Hierarchical Scoping

Energy-relevant telemetry channels are identified using semantic attribute metadata and unit normalization. To ensure operational relevance and computational efficiency, attributes with negligible economic impact are excluded. Anomaly detection is performed on the upper hierarchy levels (aggregate and floor meters), while subordinate assets are retained exclusively for diagnostic attribution.

6.4.2. Contextual Enrichment

External environmental drivers are incorporated via site-localized weather covariates derived from geographical coordinates. This contextual enrichment enables disambiguation between technical faults and demand variations induced by weather conditions.

6.4.3. Data Conditioning and Gap Resilience

Telemetry is aggregated across multiple temporal resolutions and subjected to gap-resilient conditioning. Transmission gaps and recovery artifacts are corrected through redistribution mechanisms, while integrity flags preserve explicit data quality context. Previously detected anomalous values are recursively imputed using their predicted medians to prevent contamination of subsequent inference windows.

6.4.4. Reactive Synchronization

Model inputs are refreshed on a reactive schedule aligned with finalized aggregation cycles, ensuring that inference is performed exclusively on complete and consistent telemetry segments.

6.5. Stochastic Inference and Anomaly Quantification

Anomaly detection is performed by interpreting stochastic quantile forecasts produced by Chronos-2 under a feature-driven inference strategy.

6.5.1. Feature-Driven Inference

To avoid sequential adaptation effects, inference is conditioned exclusively on historical context and exogenous covariates (weather, occupancy, temporal indicators), while the consumption value at the evaluated timestamp is withheld. This enforces contextual rather than autoregressive normative modeling.

6.5.2. Quantile-Based Scoring

Requests are batched across meters and temporal resolutions. Chronos-2 returns distribution-free quantile envelopes that define regime-conditioned normative bands. Extreme quantiles define anomaly boundaries, while central quantiles serve as conservative economic baselines.

An anomaly is triggered if the observation lies outside the extreme quantile envelope. Signed deviations are measured relative to the central quantile to quantify waste or savings under multimodal operating regimes. Deviations below a minimum economic threshold are discarded to suppress negligible events.

6.6. Hierarchical Root Cause Analysis (RCA)

Upon detection of an aggregate-level anomaly, all subordinate assets in the functional hierarchy are analyzed to localize the physical origin of the deviation.

6.6.1. Diagnostic Attribution

Sub-meter telemetry is re-evaluated to estimate asset-level financial contributions. Impacts are attributed both at individual asset level and aggregated by semantic asset categories (e.g., HVAC, lighting, plug loads), enabling identification of systematic inefficiencies and dominant fault contributors.

6.6.2. Localization and Contextual Enrichment

Attributed root causes are enriched with spatial metadata and localized environmental context to support interpretation and remediation planning.

6.7. Temporal Collapse and Persistence

Telemetry is evaluated across multiple temporal resolutions, which may generate overlapping anomaly events. A temporal collapse mechanism consolidates lower-resolution detections into higher-resolution events, preventing redundant alerts and preserving the most conservative financial impact estimate across scales.

6.8. AI Synthesis and Action Recommendation

High-impact anomalies are forwarded to an interpretation layer that synthesizes localized diagnostics, financial impact, and contextual metadata into structured human-readable explanations and remediation recommendations using a domain-specialized large language model (LLM). The LLM operates exclusively at the interpretation layer and does not influence detection or scoring logic.

6.9. Tenant-Specific Configuration

Detection sensitivity, financial thresholds, and economic parameters are governed by tenant-specific configuration profiles. These profiles enable adaptation to organization-specific risk tolerance and energy pricing while preserving strict multi-tenant isolation and auditability.

6.10. Frontend Integration for Operational Decision Support

The frontend constitutes the operational interface of the anomaly detection framework. It exposes detected anomalies, quantified financial impact, hierarchical root causes, and AI-synthesized recommendations in a multi-tenant environment, enabling human validation and remediation workflows.

6.10.1. Centralized Anomaly Registry and Validation Loop

Detected anomalies are presented in a centralized registry that supports triage, filtering, and prioritization by financial impact, severity, and affected assets (Fig. 6.2). Human operators can confirm or invalidate anomalies, forming a controlled human-in-the-loop feedback channel. Invalidated anomalies are excluded from future baseline construction to prevent contamination of normative profiles.

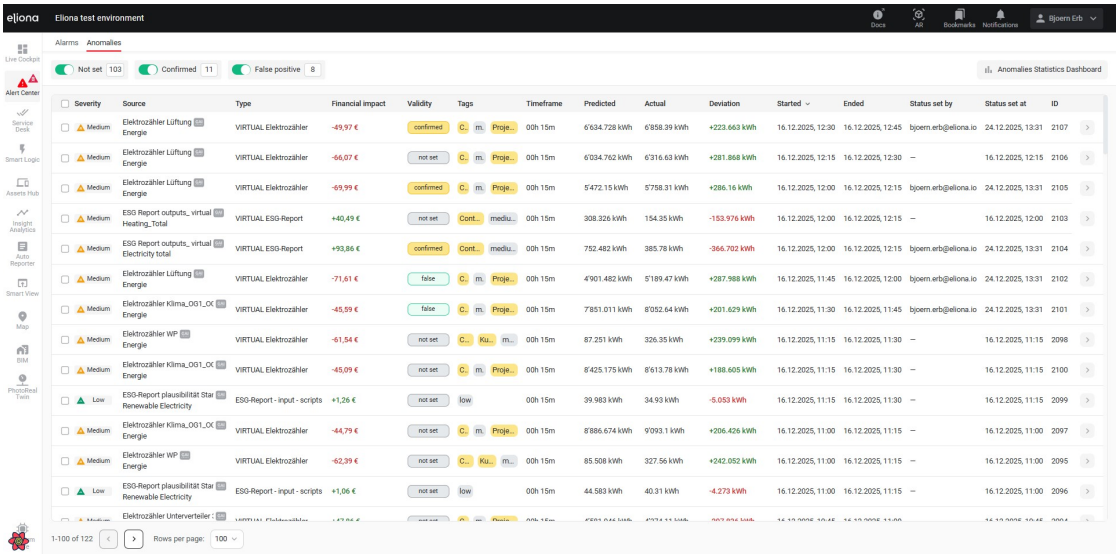


Figure 6.2: Central anomaly registry supporting prioritization, financial impact assessment, and validation workflows.

6.10.2. Quantile Visualization and Diagnostic Context

Stochastic normative envelopes and anomalous deviations are visualized directly on consumption charts using quantile-based uncertainty bands (Figs. 6.3 and 6.4). These overlays enable severity assessment relative to regime-conditioned baselines rather than absolute residual magnitudes.

6.10.3. Contextual Tooltips and AI Synthesis

Interactive tooltips expose localized financial impact, root cause attribution, validation status, and AI-generated recommendations at individual anomalous points (Fig. 6.5). This supports immediate contextual interpretation.

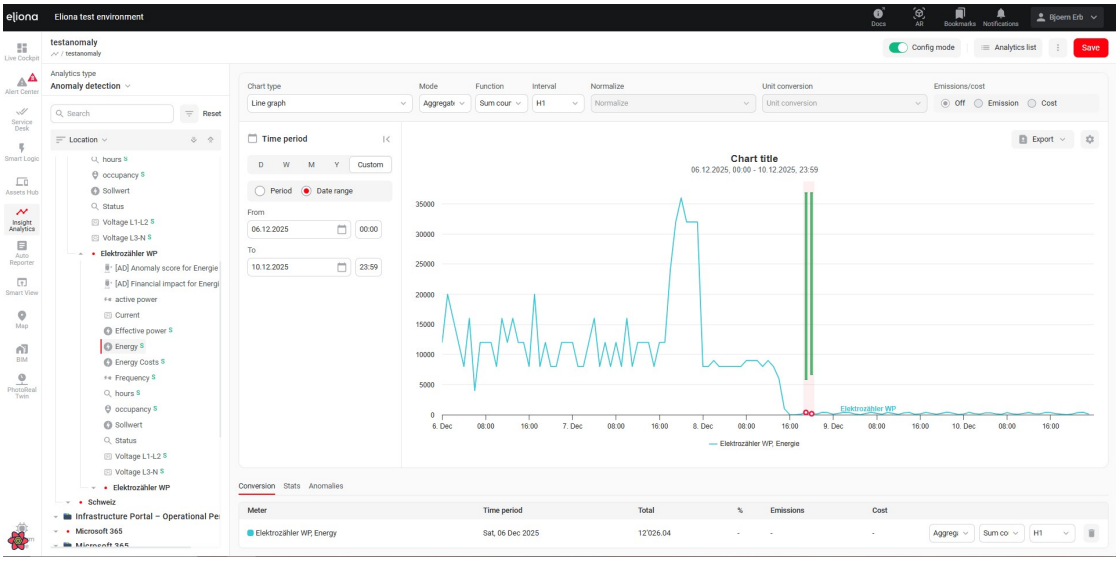


Figure 6.3: Quantile-based normative envelopes and anomaly overlays within analytics.

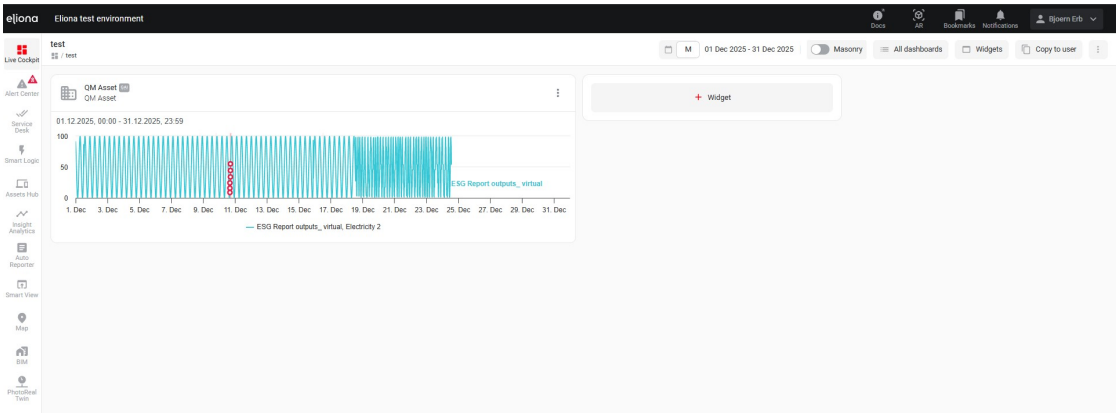


Figure 6.4: Anomaly analytics embedded into operational dashboards.

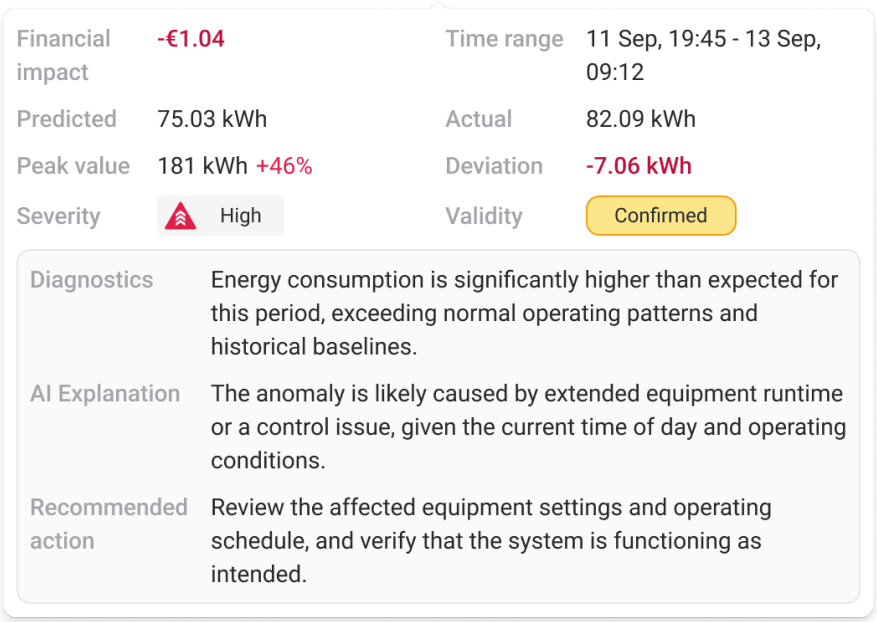


Figure 6.5: Interactive diagnostic tooltip with financial impact and AI synthesis.

6.10.4. Anomaly Detail View and Action Synthesis

A dedicated anomaly detail view consolidates quantitative metrics, hierarchical root causes, contextual metadata, and AI-generated remediation guidance into a single operational workspace (Fig. 6.6), enabling transition from detection to remediation.

6.10.5. Macro-Level Reporting and Portfolio Analytics

Portfolio-level dashboards aggregate anomalies across tenants and sites, providing executive visibility into financial impact, dominant asset categories, and temporal patterns (Fig. 6.7). Heatmap visualizations reveal systematic operational inefficiencies across time-of-day and day-of-week regimes.

6.10.6. Asset-Scoped Anomaly Integration

Anomalies are integrated into asset detail views to support localized diagnostics and historical inspection of deviations at the equipment and zone level (Fig. 6.8).

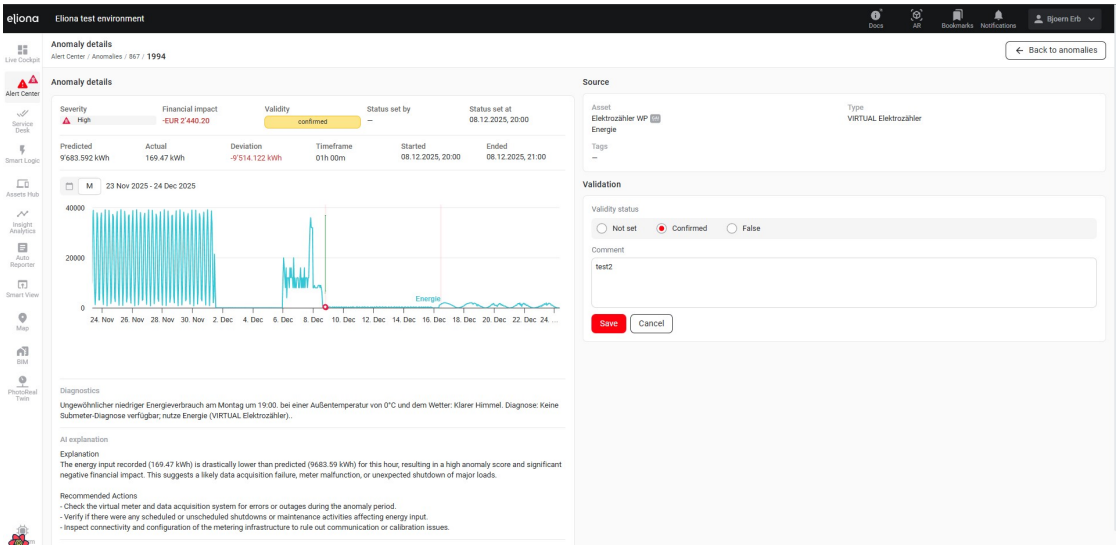


Figure 6.6: Integrated anomaly detail view combining financial impact, diagnostics, and AI-generated recommendations.

6.11. Implementation Summary

The realized system fulfills the formal design objectives by providing an infinitely scalable, cloud-native multi-tenant architecture for portfolio-scale building monitoring. It performs stochastic multivariate anomaly detection using distribution-free quantile modeling, accurately capturing multimodality and intrinsic uncertainty in non-stationary energy telemetry.

The system adapts continuously to evolving operational regimes while preventing persistent anomalies from being absorbed into normative baselines through explicit human validation mechanisms. Detected deviations are translated into conservative financial impact estimates, visualized through interpretable uncertainty envelopes, and localized via hierarchical root cause analysis. A domain-specialized large language model further synthesizes diagnostic evidence into actionable operational insights, completing an end-to-end, economically interpretable anomaly management pipeline.

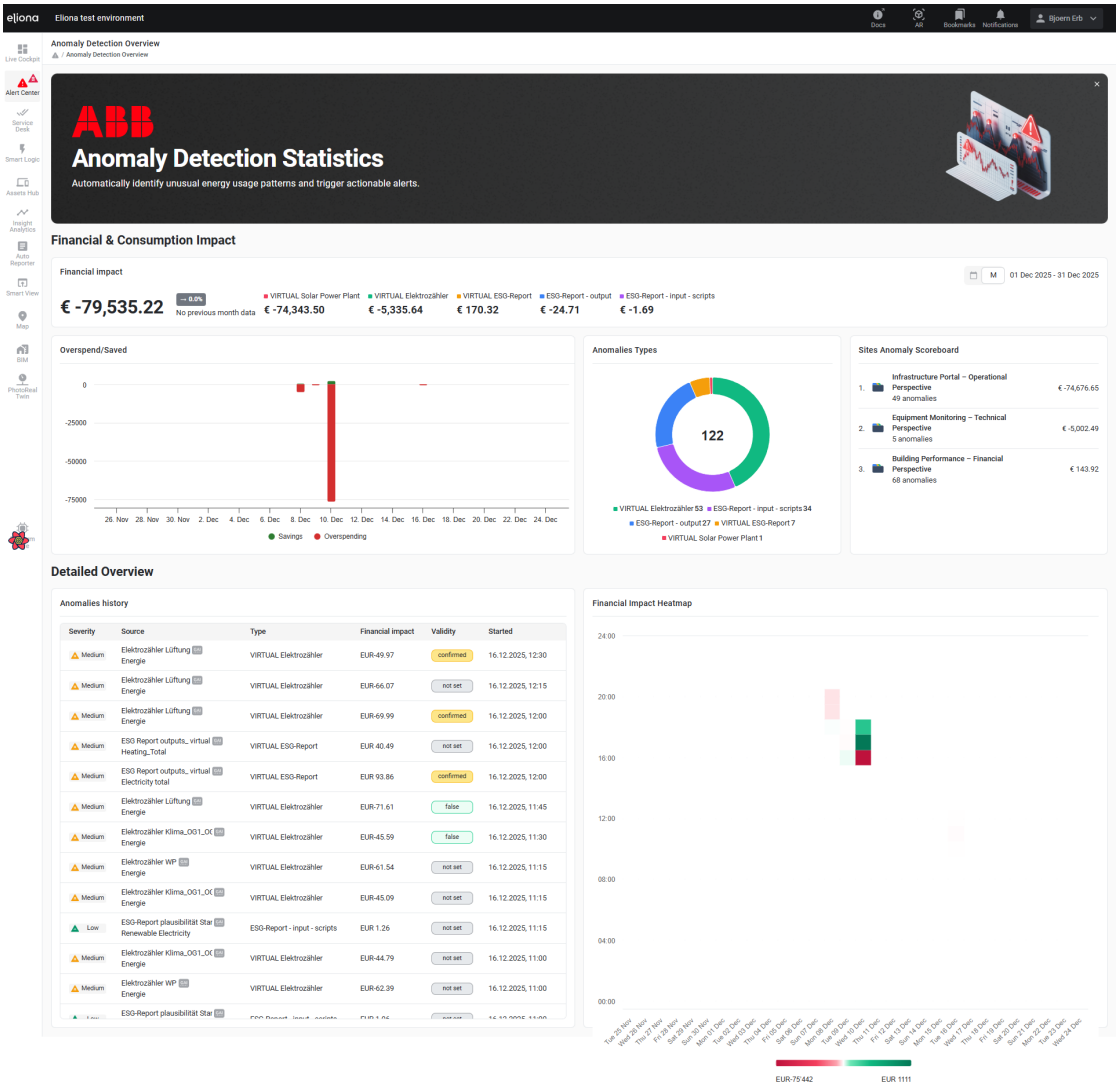


Figure 6.7: Portfolio-level anomaly analytics and financial impact aggregation.

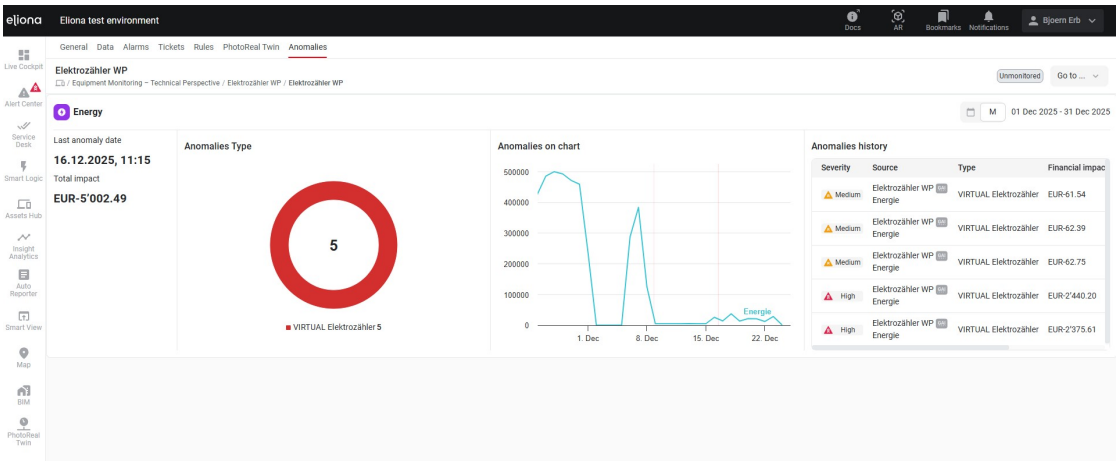


Figure 6.8: Asset-level anomaly integration enabling localized inspection and triage.

7

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.

7.1. 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 5.1) 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 CO_2 levels or access-control logs, to be included as exogenous

features for customized detection models.

7.2. 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.

7.3. 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.

7.3.1. 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.

7.3.2. 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.

7.4. 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.

8

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,

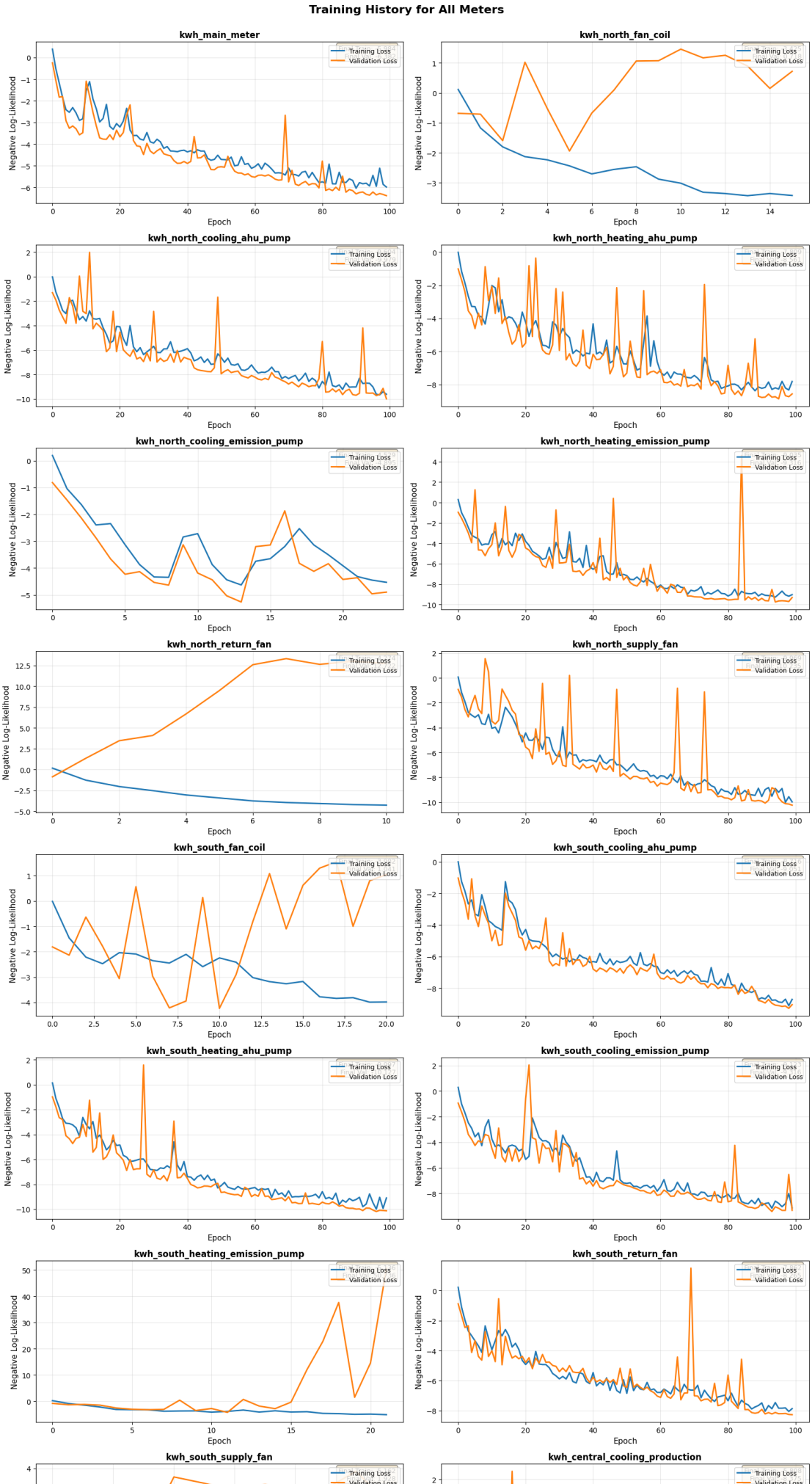
status management, and temporal heatmaps—ensures that detection events are integrated into standard maintenance workflows. The capacity for human-in-the-loop feedback, where users validate anomalies as confirmed or false, establishes a mechanism for continuous data cleansing and model refinement. This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

A

Additional Figures

A.1. Training History Across All Meters


Training History for All Meters
kwh_main_meter
kwh_north_fan_coil
kwh_north_cooling_ahu_pump
kwh_north_heating_ahu_pump
kwh_north_cooling_emission_pump
kwh_north_heating_emission_pump
kwh_north_return_fan
kwh_north_supply_fan
kwh_south_fan_coil
kwh_south_cooling_ahu_pump
kwh_south_heating_ahu_pump
kwh_south_cooling_emission_pump
kwh_south_heating_emission_pump
kwh_south_return_fan
kwh_south_supply_fan
kwh_central_cooling_production
Training Loss
Validation Loss
Negative Log-Likelihood
Epoch

References

[LGW04] Ningyun Lu, Furong Gao, and Fuli Wang. “Sub-PCA modeling and on-line monitoring strategy for batch processes”. In: *AIChE Journal* 50.1 (2004), pp. 255–259.

[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.

[LN16] Xiufeng Liu and Per Sieverts Nielsen. “Regression-based online anomaly detection for smart grid data”. In: *arXiv preprint arXiv:1606.05781* (2016).

[Peñ+16] Manuel Peña et al. “Rule-based system to detect energy efficiency anomalies in smart buildings, a data mining approach”. In: *Expert Systems with Applications* 56 (2016), pp. 242–255.

[Wu17] Jianxin Wu. “Introduction to convolutional neural networks”. In: *National Key Lab for Novel Software Technology. Nanjing University. China* 5.23 (2017), p. 495.

[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.

[Fu22] Chun Fu. *Summary of 1st Place Solution — Large-scale Energy Anomaly Detection (LEAD)*. Kaggle competition writeup. 2022. URL: https://www.kaggle.com/competitions/energy-anomaly-detection/writeups/chun-fu-summary-of-1st-place-solution (visited on 12/25/2025).

[GA22] Manoj Gulati and Pandarasamy Arjunan. “LEAD1. 0: a large-scale annotated dataset for energy anomaly detection in commercial buildings”. In: *Proceedings of the thirteenth ACM international conference on future energy systems*. 2022, pp. 485–488.

[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.

[GCM23] Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. “TimeGPT-1”. In: *arXiv preprint arXiv:2310.03589* (2023).

[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).

[Gos+24] Mononito Goswami et al. “Moment: A family of open time-series foundation models”. In: *arXiv preprint arXiv:2402.03885* (2024).

[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. 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.