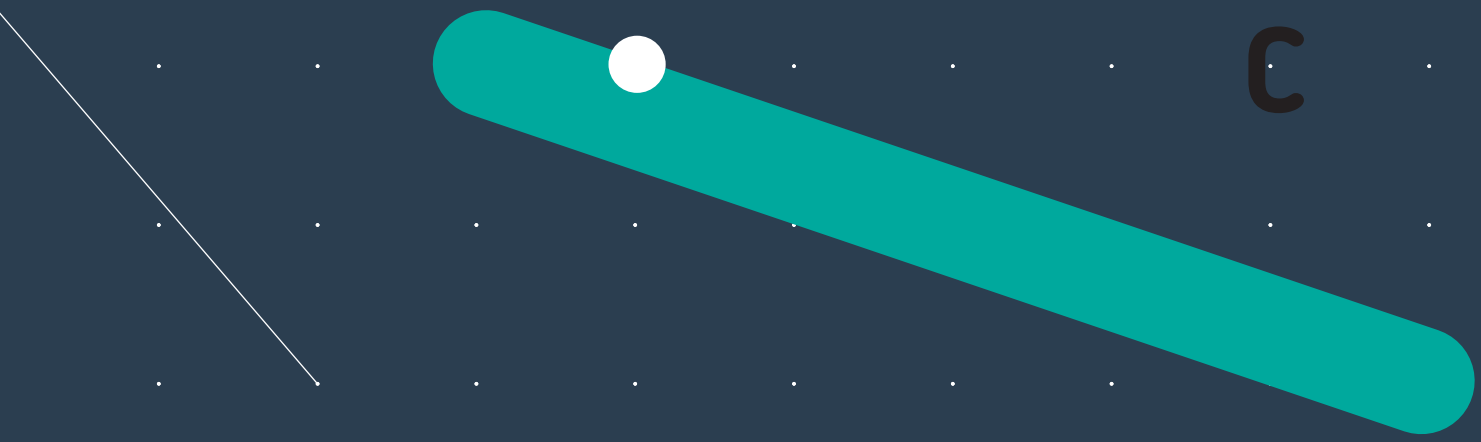
Hochschule Konstanz
Department of Computer Science

Submitted by
Samuel Tim
Student Number 307636

samuel.tim200yahoo.de



Bachelor Thesis

Towards Universal Probabilistic Foundation Models for Contextual Energy Anomaly Detection with Root Cause Attribution and Financial Impact Quantification

Konstanz, 31st December 2025

Hochschule Konstanz
Department of Computer Science

Bachelor Thesis

Towards Universal Probabilistic Foundation Models for Contextual Energy Anomaly Detection with Root Cause Attribution and Financial Impact Quantification

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 is available at https://samueltim.com/bachelor.pdf.

Abstract

Buildings account for approximately 30% of global final energy consumption. Empirical studies indicate that 4%–18% of this demand is attributable to operational anomalies such as technical faults, control and scheduling errors, and behavioural misuse, resulting in substantial avoidable energy waste.

This thesis proposes a probabilistic methodology for contextual anomaly detection in multivariate, non-stationary building-energy time series. The approach enables distribution-aware financial impact quantification and hierarchical root-cause localization and is fully realized within a production multi-tenant IoT building management platform.

A dedicated benchmark dataset is constructed using the BOPTEST simulation environment, providing clean baselines and systematically injected multivariate anomaly scenarios. Statistical, deep-learning, and foundation-model-based methods are evaluated under realistic deployment constraints.

Results demonstrate that sequential point-forecasting approaches are structurally unsuitable for anomaly detection, while stochastic probabilistic models significantly outperform deterministic predictors in multimodal operating regimes. Chronos-2 enables zero-shot portfolio-scale anomaly detection and absorbs baseline shifts without retraining. Mixture density modeling is identified as a promising foundation-model architecture, and distribution-aware quantile scoring is shown to be critical for robust anomaly quantification. 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

Acronyms

BAS Building Automation Systems

HVAC heating, ventilation and air conditioning

AMI Advanced Metering Infrastructure

IoT Internet of Things

ML machine learning

TSFM time-series foundation model

CNN convolutional neural network

RNN recurrent neural network

LSTM long short-term memory network

MDN Mixture Density Network

MD mixture distribution

PDF Probability Density Function

NLL Negative Log-Likelihood

DQ Density Quantile

QBB Quantile-Based Bounds

MSE Mean Squared Error

RP reconstruction probability

RE reconstruction error

TSAD time series anomaly detection

MCAD Multivariate Contextual Anomaly Detection

MCPA Multivariate Context Point Anomaly

RCA Root Cause Attribution

TSB-AD Time Series Benchmark for Anomaly Detection

TSB-AD-M Time Series Benchmark for Anomaly Detection - Multi-target

TSB-AD-U Time Series Benchmark for Anomaly Detection - Univariate

BOPTEST Building Optimization Performance Test Framework

LEAD 1.0 Large-scale Energy Anomaly Detection benchmark version 1.0

PA-F1 Point-Adjustment F1 score

TP true positive

TN true negative

FP false positive

FN false negative

ROC-AUC Area Under the Receiver Operating Characteristic Curve

VUS-PR Volume Under the Surface–Precision Recall

PCA principal component analysis

Sub-PCA subspace principal component analysis

LLM large language model

IPMVP International Performance Measurement and Verification Protocol

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.

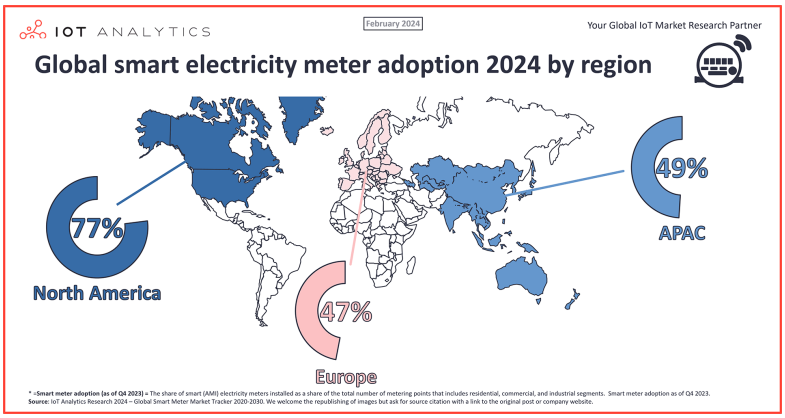


Figure 1.1: Global smart electricity meter adoption by region in 2024, illustrating the varying levels of AMI penetration across markets [IoT24].

1.3. Why Current Building Automation Systems Fail

Most deployed 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, particularly the sequential-forecasting failure modes analyzed in Section 4.6. Many approaches operate on deterministic point predictions that fail to represent the stochastic, multimodal, and non-stationary characteristics detailed in Section 2.1 and empirically dissected in Sections 4.7 and 4.8. 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, as demonstrated in Section 4.7.

Sequential forecasting-based detectors further suffer from two critical failure modes

when applied to sustained anomalies:

1. error propagation, where anomalous values corrupt the sliding input window and destabilize future predictions, as analyzed in Section 4.6.2;

2. rapid baseline adaptation, where models quickly absorb anomalous states as normal operation, causing long-duration anomalies to disappear from the anomaly score signal, as detailed in Section 4.6.3.

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, as detailed in Section 3.2. 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

extual deviations to be distinguisheextual deviations to be distinguisheThe 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 distinguish contextual deviations from normal variability. The integration of TSFM into the proposed pipeline is detailed in Section 4.3.

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 Smart Building Management Platform [Eli25a] (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 under large-scale, non-stationary operating conditions.

In contrast to threshold-based monitoring and deterministic forecasting-based detectors, anomaly detection is formulated as a stochastic, multivariate, and 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 quantifies their financial impact, localizes probable root causes within building hierarchies, and synthesizes operationally meaningful recommendations.

Methodologically, the work targets Multivariate Context Point Anomaly (MCPA) under multimodality and non-stationarity, thereby enabling probabilistic deviation semantics and economically interpretable scoring (see Sections 4.7 and 4.8).

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 (see Sections 2.1 and 4.7).
- An empirical and theoretical critique of deterministic sequential forecasting-based

anomaly detectors and their structural failure modes (see Section 4.6 and Sections 4.6.2–4.6.3).

- A stochastic detection framework based on probabilistic normative bands derived from time-series foundation models, enabling contextual anomaly scoring (see Section 4.3).
- A distribution-aware anomaly scoring formulation for mixture-density outputs, including Quantile-Based Bounds (QBB) and Density Quantile (DQ) based scoring and severity scaling for robust quantification under multimodality (see Section 4.8).
- A financially interpretable quantification layer that transforms deviations into conservative monetary impact estimates (see Section 4.4).
- A hierarchical root-cause attribution pipeline grounded in building ontologies, including aggregation by asset type and action synthesis (see Section 4.5).
- A domain-specific multivariate benchmark dataset generated via Building Optimization Performance Test Framework (BOPTEST) with labeled contextual fault scenarios and controlled seasonal segmentation (see Section 5.1).
- A critical evaluation protocol with explicit discussion of training stability, coverage bias, and comparability constraints between trainable models and foundation models (see Section 5.4).
- A fully integrated, scalable, multi-tenant implementation deployed within a production smart building Internet of Things (IoT) platform, including human-in-the-loop validation mechanisms to prevent baseline contamination and multi-resolution persistence handling (see Chapter 6).
- A methodological basis for standardized, International Performance Measurement and Verification Protocol (IPMVP)-aligned baseline comparison and future universal energy foundation models supporting zero-shot anomaly detection at portfolio scale (see Section 7).

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.

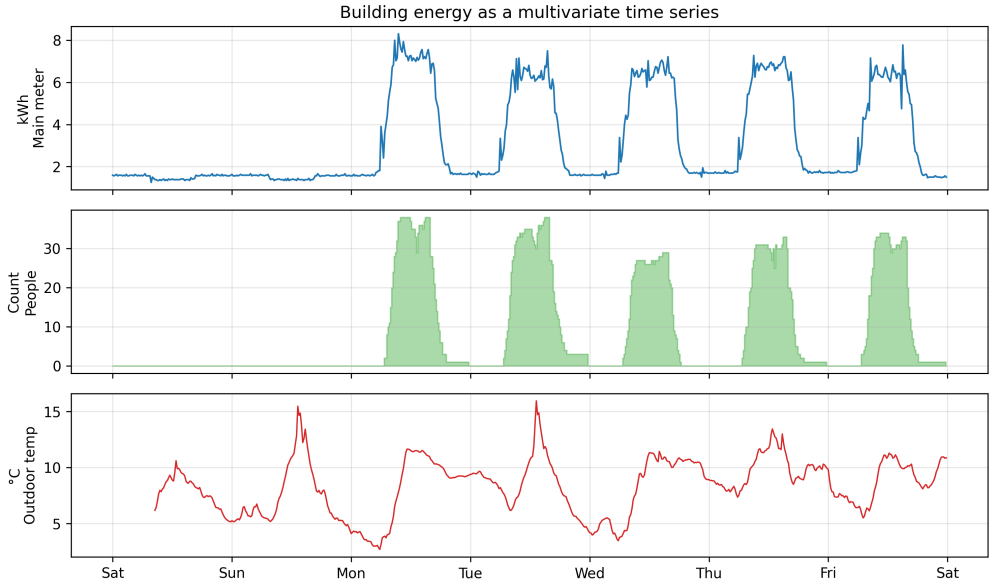


Figure 2.1: Representative multivariate time series showing the main meter load together with occupancy (people count) and outdoor temperature. The plot illustrates how multiple interdependent variables evolve jointly over time.

2.1.2. Causal Chain of Energy Consumption

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 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.

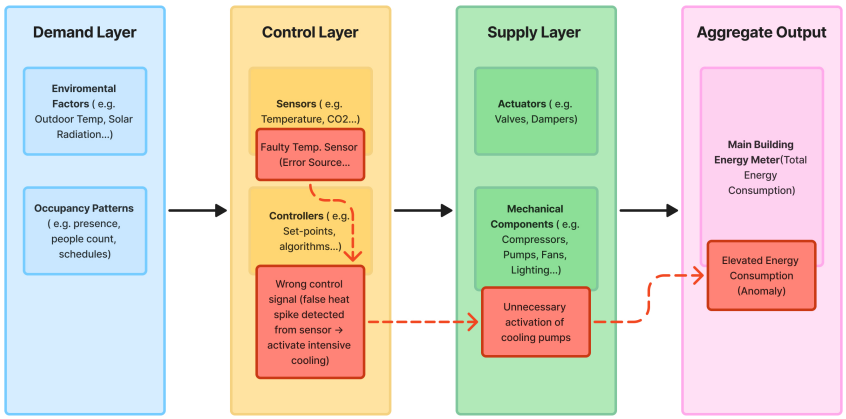


Figure 2.2: Causal chain of building energy consumption from demand over control to supply layer.

HVAC and Environmental Drivers

heating, ventilation and air conditioning (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 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.

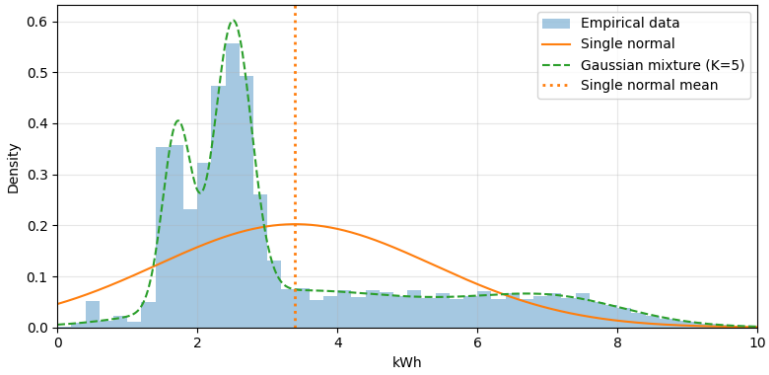


Figure 2.3: Empirical distribution of the building's main meter (15-minute kWh values, histogram) with an overlaid single normal distribution and a Gaussian mixture model with five components, illustrating the mismatch between a unimodal Gaussian model and the multimodal, heavy-tailed structure of real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

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 [Eli25d]. 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.

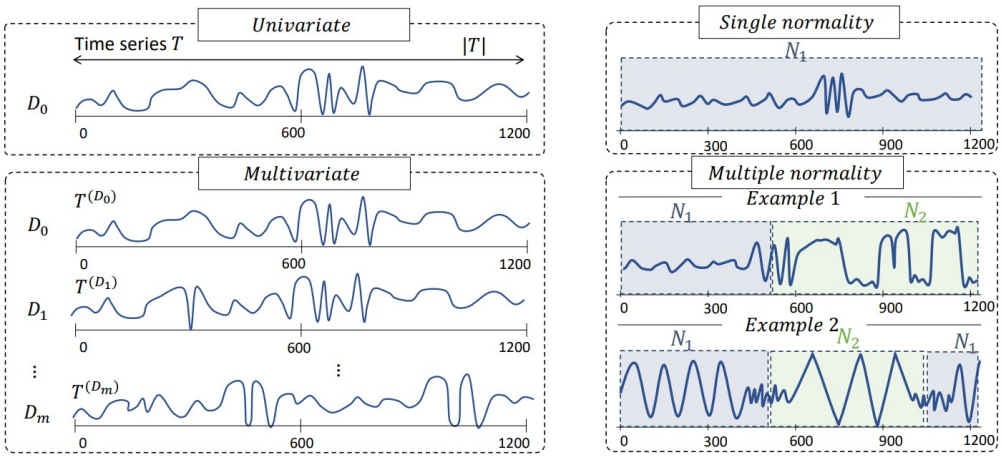


Figure 2.4: Schematic illustration of time series types along two axes—dimensionality (univariate vs. multivariate) and normality regimes (single-mode vs. multi-mode). Adapted from Boniol et al.'s tutorial on new trends in time series anomaly detection [BPP23].

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

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

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

2.2.2. 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.

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.

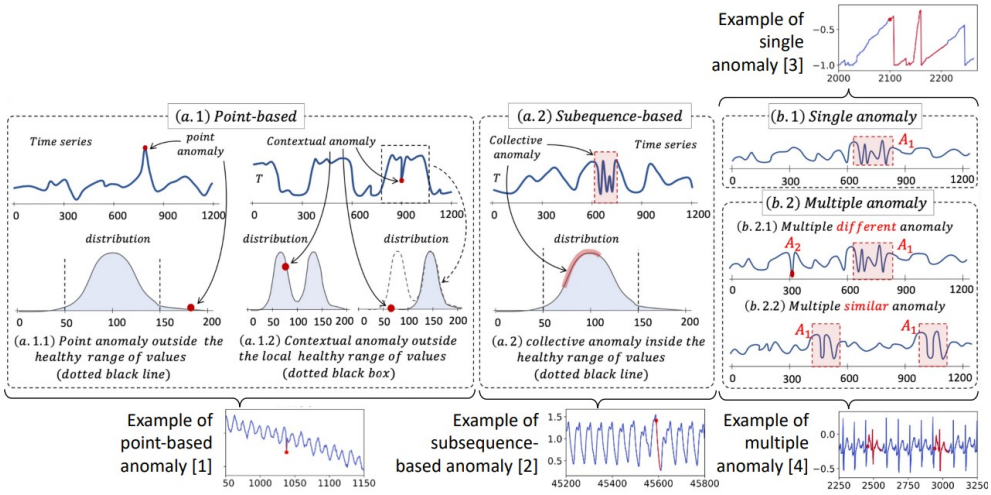


Figure 2.5: Taxonomy of time series anomalies along structural and multiplicity dimensions, distinguishing global and contextual point anomalies, subsequence-based anomalies, and their occurrence as single, multiple different, or multiple similar events. Adapted from Boniol et al.'s tutorial on new trends in time series anomaly detection [BPP23].

2.3. Methodological Approaches to Anomaly Detection

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

2.3.1. Anomaly Scores

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

2.3.2. Learning Paradigms

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

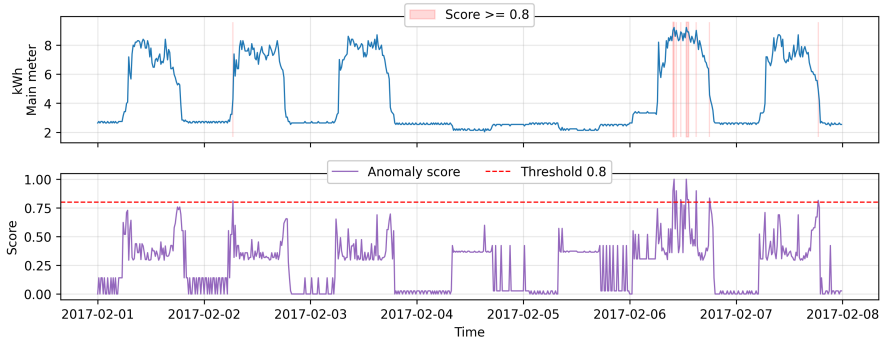


Figure 2.6: Example of an anomaly score $s_i \in [0, 1]$ aligned with the underlying time series. A threshold of 0.8 separates normal points from those flagged as anomalous.

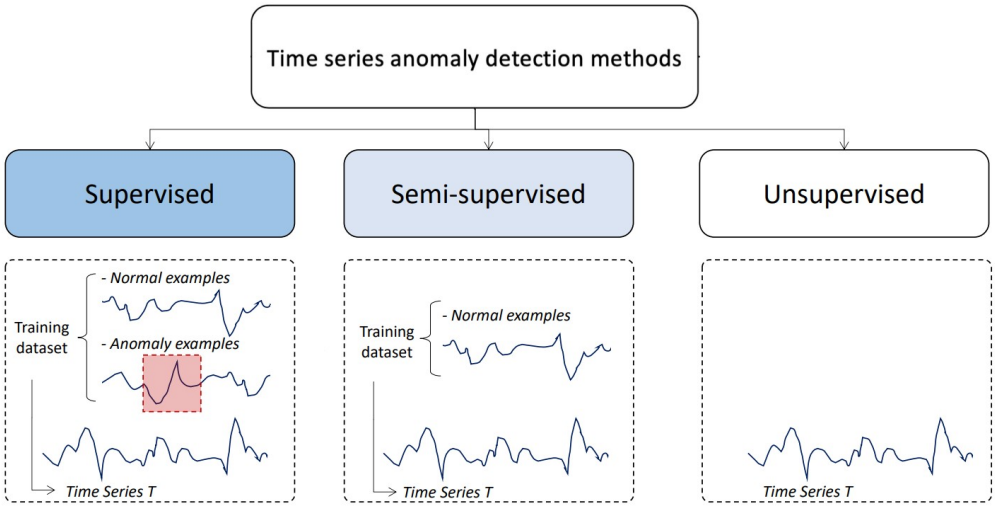


Figure 2.7: Schematic overview of supervised, semi-supervised, and unsupervised learning paradigms for anomaly detection, adapted from Boniol et al.'s tutorial on new trends in time series anomaly detection [BPP23].

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.

In building-energy monitoring, the practical setting is predominantly semi-supervised: historical telemetry is typically available and is treated as mostly nominal baseline behavior, while explicit anomaly labels are sparse or absent.

2.3.3. Families of Detection Methods

Anomaly detection approaches are grouped into three methodological families:

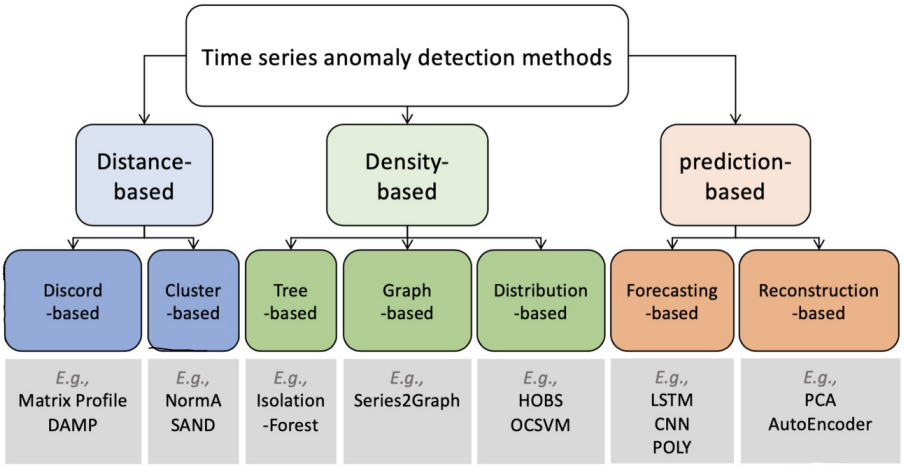


Figure 2.8: Hierarchical taxonomy of anomaly detection methods grouped into distance-based, density-based, and prediction-based families, adapted from Boniol et al.'s tutorial on new trends in time series anomaly detection [BPP23].

Distance-Based: Identify anomalous subsequences by pattern dissimilarity.

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

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

This thesis focuses on prediction-based approaches, as they are the only methodological family that provides an explicit expected-value baseline, enabling deviations to be quantified in physical units and directly translated into financial impact. This property is essential for contextual building-energy anomaly detection and economic loss estimation.

2.4. Benchmarking Foundations

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 positive (TP), true negative (TN), false positive (FP), and false negative (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 is a multivariate, multimodal, and non-stationary stochastic process governed by a causal chain spanning demand, control, and mechanical execution (see Sections 2.1.1–2.1.2). Consequently, deviations at aggregate meters typically reflect upstream technical or behavioural faults rather than isolated numerical outliers.

From these structural and statistical properties, the following requirements for anomaly detection follow (see Sections 2.1.1, 2.1.3, and 2.1.7):

1. **Probabilistic multivariate modeling**
   Expected behaviour should be represented probabilistically to capture multimodal operating regimes and cross-variable dependencies (see Section 2.1.5).

2. **Robustness to temporal dependence and persistence**
   Detection must remain stable under autocorrelation, seasonality, and long-duration persistence such that sustained deviations are not absorbed into the baseline (see Section 2.1.3).

3. **Separation of physical faults and digital artifacts**
   Transmission gaps and aggregation artifacts must be separated from physical anomalies through explicit data-quality handling (see Section 2.1.7).

Building-energy telemetry is dominated by contextual point anomalies at aggregation horizons relevant for monitoring (see Sections 2.2.1 and 2.2.3). Persistent deviations already present in historical baselines may be absorbed into learned normality (normality drift), representing a fundamental constraint of semi-supervised and unsupervised 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.

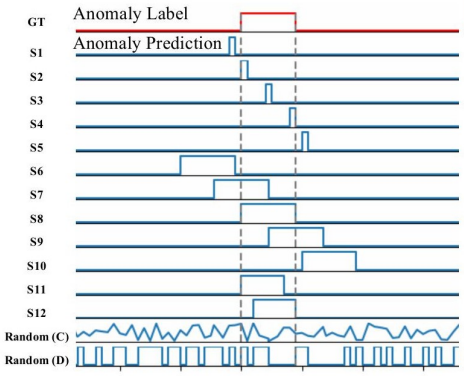
Regression-based baselining methods constitute another classical detection paradigm. Liu and Nielsen [LN16] propose an online regression framework for smart-meter anomaly detection in which expected consumption is estimated via supervised learning models and anomalies are detected as residual deviations. 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 TSAD is often an artifact of flawed evaluation practices rather than algorithmic superiority.

3.2.1. Systemic Flaws and Metric Reliability



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].

Historical results are often compromised by three documented data-level flaws. First, mislabeling leads to artificially high FN 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

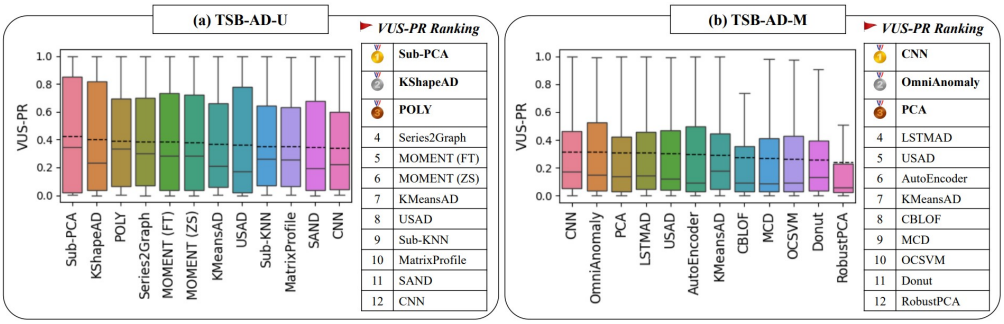


Figure 3.2: Accuracy evaluation of the top 12 methods on (a) univariate Time Series Benchmark for Anomaly Detection - Univariate (TSB-AD-U) and (b) multi-target Time Series Benchmark for Anomaly Detection - Multi-target (TSB-AD-M) datasets based on the VUS-PR metric. Adapted from Liu and Paparrizos [LP24].

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 multi-target scenarios (TSB-AD-M). As shown in Figure 3.2, convolutional neural networks (convolutional neural network (CNN)) [Wu17] and generative models like OmniAnomaly [Su+19] consistently outperform statistical baselines in capturing non-linear dependencies across multiple sensor channels [Su+19].

3.2.3. Implications for Multivariate Context Point Anomalies

While the Time Series Benchmark for Anomaly Detection (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 multi-target dependency modeling, while TSFM demonstrate superior efficacy in point anomaly identification. However, the TSB-AD-M partition contains a limited representation of multi-target 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 TSFM in univariate con-

texts, 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 (MCAD) 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 MCAD 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 Large-scale Energy Anomaly Detection benchmark version 1.0 (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 Area Under the Receiver Operating Characteristic Curve (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 1.0 effectively learn to imitate human visual judgments rather than to detect physically or economically relevant inefficiencies. These properties limit the transferability of LEAD 1.0-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, TSFM. 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 [Su+19], 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

TSFM 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 TSFM in forecasting-residual-based anomaly detection tasks. Specifically, variational autoencoders trained from scratch surpass large TSFM 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, TSFM capable of native distributional forecasting—such as Chronos-2 [Ans+25]—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. 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 because it best satisfies the thesis objectives under portfolio-scale constraints. In particular, the zero-shot, context-conditioned operating mode of TSFMs directly supports **O3** by adapting to baseline drift and abrupt regime changes through recent-context conditioning rather than scheduled per-meter retraining pipelines. This simultaneously strengthens practical scalability (**O7**) and preserves context-conditioned normality (**O1**).

Trainable baselines require explicit per-meter fitting and repeated retraining to track non-stationarity, implying fragile orchestration at scale. Persisting anomalies are prevented from contaminating the adaptive reference by applying Strategy B from Section 4.6.4, which operationalizes **O4** in the deployed pipeline.

While Chronos-2 does not expose an explicit multimodal mixture density, it provides calibrated, distribution-free quantile bounds (Section 4.8.2), which satisfy **O2** for deployment. These quantile envelopes enable QBB anomaly scoring without parametric assumptions and avoid the instability and coverage gaps observed for mixture-density architectures, while remaining competitive in detection accuracy.

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 Smart Building Platform

The anomaly detection system is integrated into the Eliona smart building platform, which serves as the operational environment for data ingestion, storage, and visualization[Eli25a]. 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

Figure 4.1 summarizes the platform architecture. Eliona is a multi-tenant system in which multiple building portfolios share infrastructure while tenant data and configurations remain logically isolated.

The stack comprises three layers:

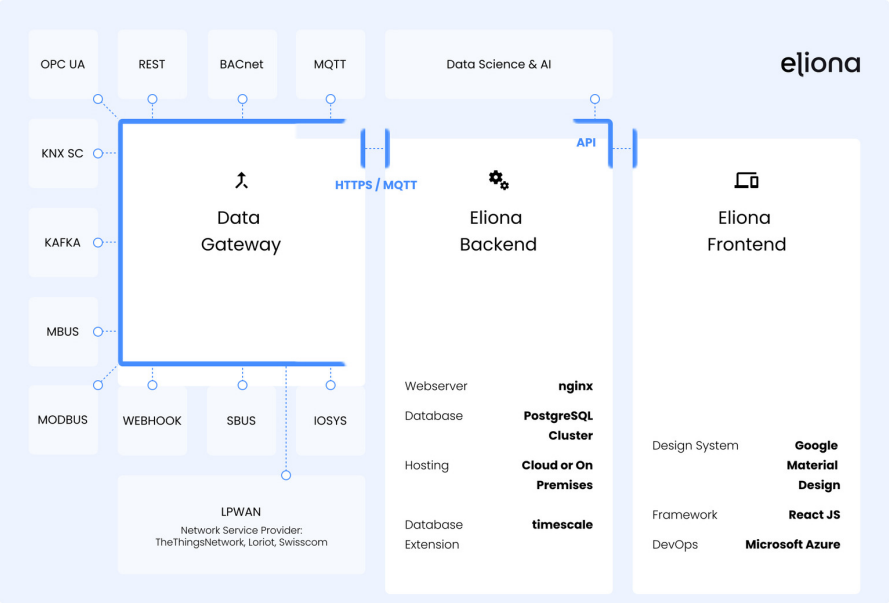


Figure 4.1: Layered architecture of the Eliona platform, from protocol ingestion to analytics and user interaction. Source: Eliona documentation[Eli25c].

Data Gateway Terminates heterogeneous building and IoT protocols (e.g., BACnet, MQTT, OPC UA, REST, Modbus) and forwards normalized telemetry to the platform via secure transport (HTTPS/MQTT).

Backend Provides the central service and persistence layer, implemented as a set of microservices for tenant and asset management, authentication and authorization, ingestion, rule-based automation, and analytics orchestration[Eli25e; Eli25f]. Data are persisted in a shared database with strict tenant scoping and exposed through a unified API.

Frontend Consumes this API and provides role-aware user workflows for portfolio navigation, asset-hierarchy exploration, dashboarding and digital twins, and investigation of alarms and anomalies and much more.

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[Eli25d; Eli25b]. 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 → Building → Floor), while the Functional Tree captures technical dependencies (e.g., Heating System → Pump → 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 following objectives are derived from the structural properties of building-energy telemetry and the methodological constraints identified in prior work (see Chapters 2 and 3). They define the design targets that govern both the detection methodology and its production deployment.

4.2.1. Methodological Objectives

O1: Contextual Fidelity
Deviations must be assessed relative to a context-conditioned normative baseline (seasonality, occupancy, weather) rather than absolute residual magnitude, to prevent regime changes from being misclassified as anomalies.

O2: Distributional Validity
Normal behavior must be represented by calibrated predictive uncertainty (e.g., quantile envelopes) that remains valid under multimodal, heavy-tailed, and non-Gaussian operating regimes.

O3: Drift Robustness
Detection must remain sensitive under gradual baseline drift and abrupt regime transitions (baseline shifts) without relying on fragile, scheduled retraining cycles.

O4: Persistence Preservation
Sustained faults and recurring inefficiencies must remain detectable and must not

be absorbed into the normative reference through adaptation effects (e.g., sliding-window contamination or implicit normalization).

O5: Hierarchical Diagnosability
Detected anomalies must be localizable to concrete physical subsystems via hierarchical asset ontologies to support actionable diagnostics.

O6: Economic Interpretability
Deviations must be translated into conservative, physically defensible signed impact estimates (loss or savings) based on a regime-consistent baseline.

4.2.2. System and Deployment Objectives

O7: Horizontal Scalability
The framework must support horizontally scalable, parallel processing of large building portfolios with independent execution across assets and tenants.

O8: Multi-Tenant Isolation
The system must enforce strict tenant scoping for data, configuration, and outputs, and prevent cross-tenant interference (both logical leakage and resource contention).

O9: Data Integrity Robustness
Transmission gaps, aggregation spikes, and recovery artifacts must be explicitly handled to avoid generating non-physical anomalies.

4.3. Time-Series Foundation Models

TSFMs directly support Objectives **O2**, **O3**,**O1** and **O7** by providing probabilistic forecasts with context-conditioned adaptation under non-stationarity at portfolio scale, reducing the need for per-asset retraining pipelines. TSFMs are large-scale pretrained sequence models trained on heterogeneous collections of time series across domains, tasks, and temporal resolutions. In contrast to task-specific forecasting networks, TSFMs learn general temporal representations and can be applied in a zero-shot regime to unseen series, removing the need for maintaining and periodically retraining asset-specific models.

Chronos-2 is a transformer-based TSFM trained on large-scale synthetic and real-world corpora and provides direct probabilistic quantile forecasts [Ans+25]. By condi-

tioning inference on the most recent historical context (and optional covariates), Chronos-2 can track baseline shifts through its context window rather than explicit retraining. This combination of calibrated distributional outputs and context-based adaptation makes it well aligned with non-stationary building-energy telemetry and large-portfolio deployment constraints.

4.4. Financial Impact Quantification

This section addresses Objectives **O6** and **O2** by defining a signed, regime-consistent impact estimator that maps deviations to conservative monetary loss or savings.

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.

4.4.1. 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 with maximal posterior responsibility for x_t (MAP 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.

4.4.2. Fallback Without Mixture Information

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

$$\tilde{\mu}_t = \mu_t + \operatorname{sign}(x_t - \mu_t)\,\sigma_t.$$

4.4.3. 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.

4.5. Hierarchical Root Cause Analysis and Action Synthesis

This section addresses Objective **O5** by localizing aggregate deviations to concrete subsystems using the asset ontology and summarizing evidence for operational interpretation.

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

4.5.1. Ontology-Guided Attribution

Assets are organized in dual hierarchies capturing geographical containment (e.g., site, building, floor) and functional decomposition (e.g., main meter, sub-meter, device). When an anomaly is detected at an aggregate level, subordinate assets are queried and ranked by their cumulative signed impact over the anomaly interval. The highest-ranked assets are treated as the primary contributors to the aggregate deviation, enabling localization to specific meters, subsystems, and physical zones.

4.5.2. Aggregation by Asset Type

To expose distributed inefficiencies, impacts are additionally aggregated by semantic asset categories (e.g., HVAC, lighting, plug loads). This highlights systematic issues that may be small per device but substantial in combination.

4.5.3. Contextual Synthesis and Action Generation

Localized contributors, type-level aggregates, and contextual conditions (time, calendar, weather) are consolidated into a structured diagnostic payload and passed to a domain-specialized large language model (LLM). The LLM operates exclusively at the interpretation layer, converting structured evidence into human-readable explanations and actionable recommendations, including quantified impact estimates.

4.5.4. Design Rationale

Detection, scoring, and impact quantification are derived deterministically from measured data and the asset ontology. The LLM is restricted to semantic interpretation and does not influence anomaly detection or decision boundaries, ensuring traceability and auditability.

4.6. Critique of Sequential Forecasting for Anomaly Detection

This section addresses Objectives **O4** and **O3** by demonstrating how autoregressive detectors can become unstable or adapt away sustained anomalies under non-stationarity.

A dominant paradigm in time-series anomaly detection is the use of sequential forecasting models. In this approach, a model (e.g., recurrent neural network (RNN), long short-term memory network (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.

4.6.1. 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.

4.6.2. Failure Mode 1: Error Propagation and Instability

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

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

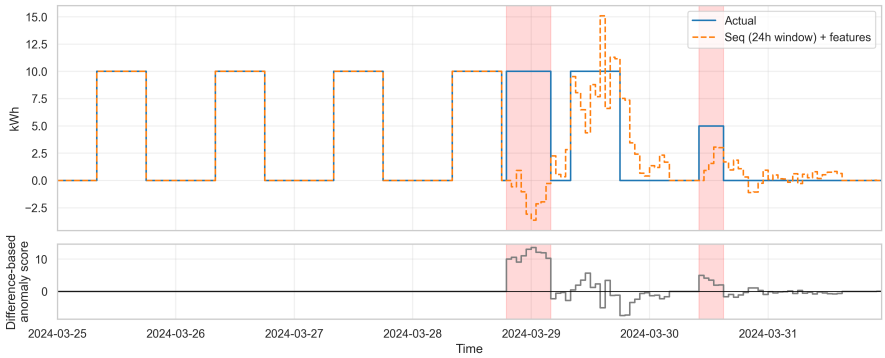


Figure 4.2: Prediction behavior of a model using a 24 h historical window plus time features. The top panel shows actual vs. predicted values; the bottom panel shows the difference-based anomaly score. Note the erratic predictions even after the anomaly ends as the unseen data propagates through the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

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.

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

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

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

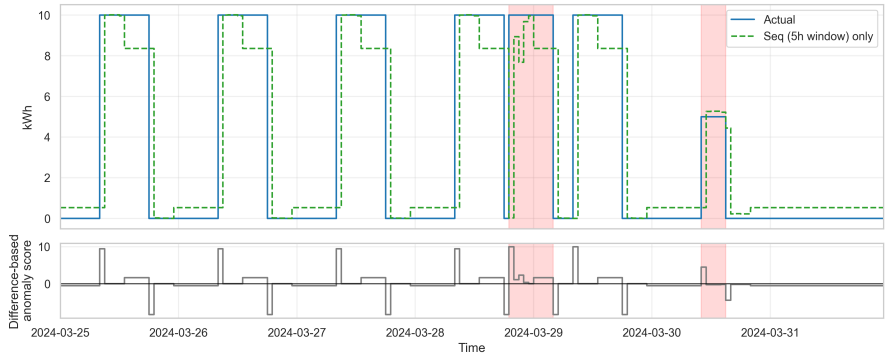


Figure 4.3: Prediction behavior of a model using only a short (5-step) historical window. The model correctly identifies the onset of anomalies but rapidly adapts to the new level, causing the anomaly score to drop back to near zero while the anomaly is still ongoing.

The model successfully flags the onset of both anomalies due to the sudden jump. However, within five time steps, the input window is filled with the anomalous values. The model quickly learns the new "normal" (e.g., that consumption is currently 10 at night) and predicts accordingly. The residual drops to near zero, and the anomaly is effectively missed for the majority of its duration.

Implications for Evaluation Metrics and Financial Impact

This behavior explains the heavy reliance in academic literature on 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.

4.6.4. Mitigation Strategies

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

Strategy A: Contextual Feature-Only Modeling

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

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

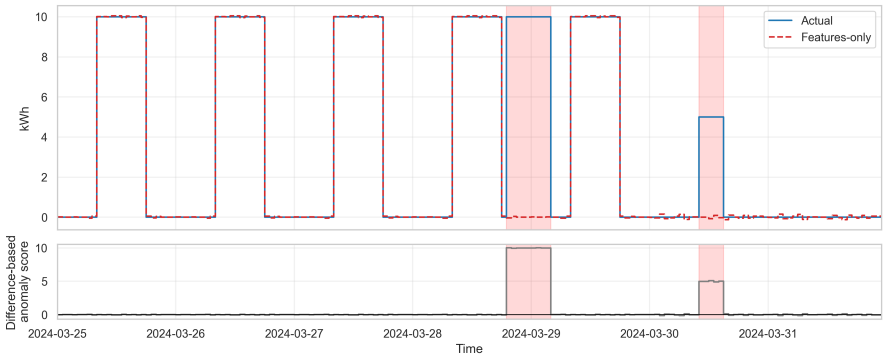


Figure 4.4: Behavior of a features-only model (no historical consumption input). The prediction relies solely on context (time/weekend), resulting in a stable baseline and accurate detection of sustained anomalies without adaptation.

Strategy B: Inference-Time Input Imputation

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

events are not reinterpreted as normal behaviour despite the non-stationarity of the raw building signal.

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

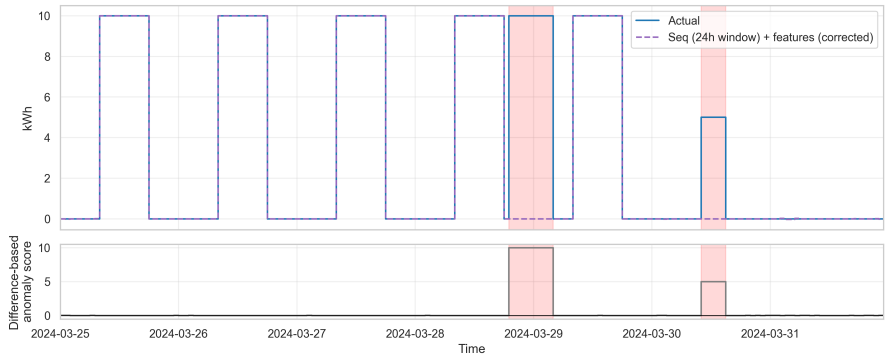


Figure 4.5: The same 24-hour window model from Figure 4.2, but applied with inference-time imputation. When an anomaly is detected, the predicted value replaces the actual value in the sliding window for future steps. This prevents error propagation and maintains a high anomaly score throughout the event.

4.7. Statistical Limitations of Point and Gaussian Predictions

This section addresses Objective **O2** by demonstrating why point and single-Gaussian predictors fail under multimodal regimes.

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

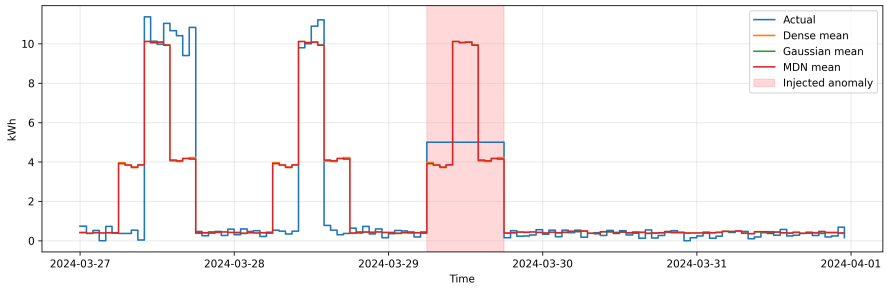


Figure 4.6: Predicted means over the Variable Shift horizon. Dense, Gaussian, and MDN models track the two regimes, masking the scoring deficiencies discussed in Sections 4.7.1–4.7.3.

4.7.1. 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 $5\,\text{kWh}$ regardless of whether the system is in its “Off” (low) or “On” (high) state. Consequently, perfectly normal behavior is scored as highly anomalous, whereas the injected stuck-at-5 event appears deceptively healthy because it matches the biased mean. The residual trace in Figure 4.7 exposes this contradiction: the absolute error balloons whenever the device operates normally, yet it contracts when the genuine anomaly occurs.

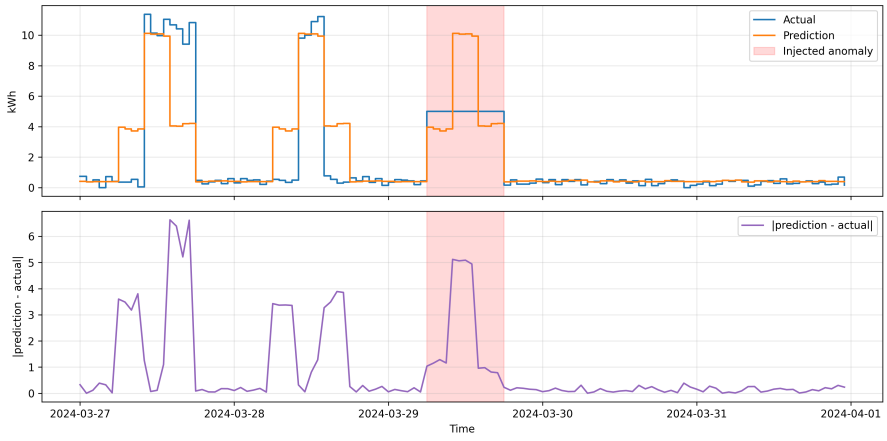


Figure 4.7: Dense regressor residuals over the Variable Shift dataset. The mid-range prediction inflates anomaly scores for legitimate operating states, while the stuck-at-5 fault yields a small residual.

4.7.2. 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 Negative Log-Likelihood (NLL) trace (Figure 4.8) confirms that the stuck-at-5 anomaly sits inside the "most likely" area of the Gaussian, generating a low penalty. Meanwhile, legitimate regime values land in lower-density shoulders and spuriously raise the score. The heatmap in Figure 4.9 makes the distortion visible: the green, high-probability band spans the median instead of the true modes.

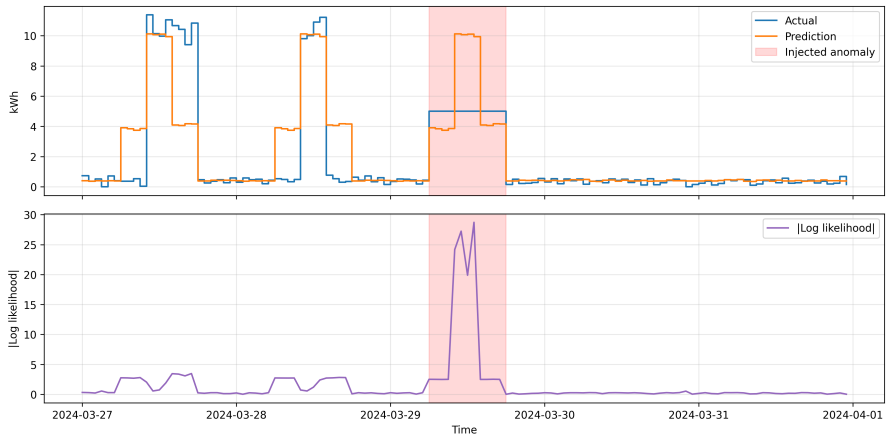


Figure 4.8: Absolute log-likelihood trace for the single-Gaussian predictor. The stuck-at-5 anomaly aligns with the high-likelihood center, suppressing the score.

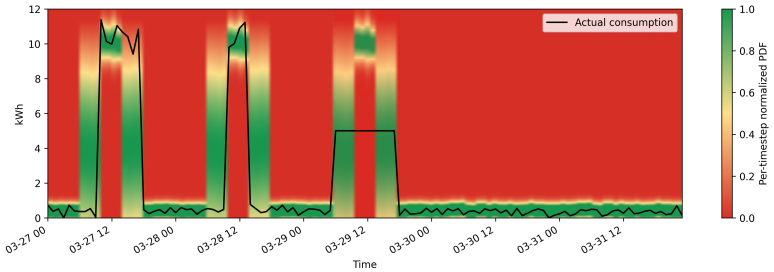


Figure 4.9: Per-timestep normalized PDF for the Gaussian model. High probability mass accumulates between the actual clusters, illustrating the variance-stretching paradox.

4.7.3. 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 re-

gions between components retain near-zero probability. Figure 4.11 shows how the MDN assigns green (high probability) bands only where data is observed, keeping the mid-range red. When log-likelihood is used as the anomaly score, the stuck-at-5 fault immediately falls into the valley between components, producing a sharp increase in $|\log p(x)|$ (Figure 4.10). This probabilistic separation allows the MDN to quantify financial impact reliably: integrating the residual energy over time now reflects the true magnitude of the fault rather than artifacts of model bias.

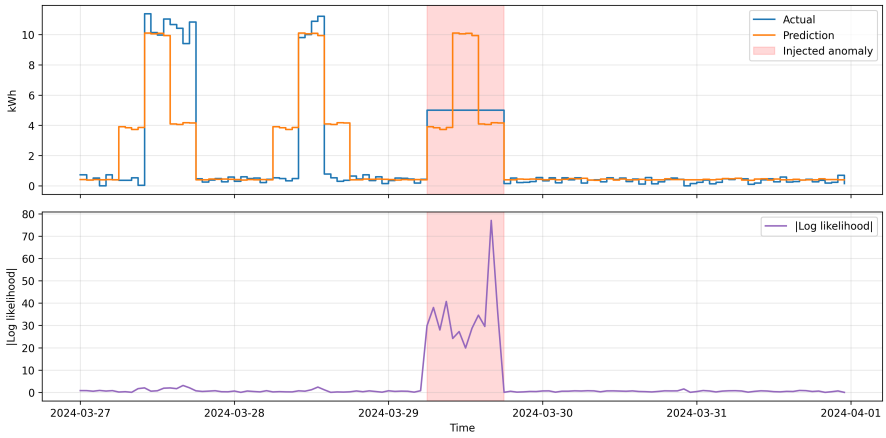


Figure 4.10: MDN absolute log-likelihood trace. The stuck-at-5 anomaly triggers a sustained spike because the value resides in a low-probability region between mixture components.

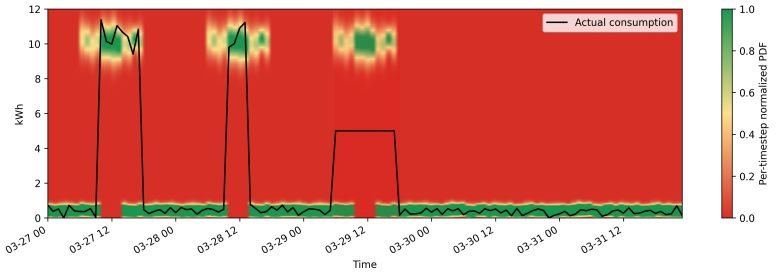


Figure 4.11: MDN normalized PDF heatmap. Two distinct high-probability ridges align with the real operating modes, while the middle band remains improbable.

4.8. Distribution-Aware Anomaly Scoring for Mixture Distributions

This section addresses Objectives **O2** and **O4** by deriving scores that remain valid under multimodality and are comparable across time for persistence tracking and ranking.

Building on Section 4.7, 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 4.6 are amplified in this setting because no single "expected value" exists, making deviation from the mean a misleading proxy for abnormality.

Figure 4.12 extends the mixture distribution (MD) perspective by overlaying several candidate scores derived from the same mixture distribution: mean residuals, PIT, negative log-likelihood, and the proposed DQ family.

4.8.1. 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 4.12, the MD 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.

4.8.2. Quantile-Based Bounds

Instead of using the mean, anomalies can be detected using outer QBB of the predictive distribution. Let $q_{t,\alpha}$ denote the α-quantile of $p_t(x)$. For a chosen α, the nominal band is $[q_{t,\alpha},\, q_{t,1-\alpha}]$. An observation is flagged if it exceeds this envelope:

$$x_t > q_{t,1-\alpha} \quad \text{or} \quad x_t < q_{t,\alpha}.$$

QBB are distribution-free and remain meaningful under non-Gaussian and heavy-tailed regimes, but they primarily detect tail events and can miss low-density valleys between modes.

Figure 4.12: Comparison of anomaly scoring methods derived from a mixture distribution (MD) under a bimodal operating regime with an injected intermediate anomaly. The top panel shows the predictive density together with the actual observation and the MD mean. Subsequent panels compare mean residuals, Quantile Based scores, NLL, and the proposed DQ scores and severities.

4.8.3. Negative Log-Likelihood and Its Limitations

Another principled score is the NLL:

$$s_t^{\text{NLL}} = -\log p_t(x_t). \tag{4.2}$$

NLL correctly assigns high anomaly scores to observations in low-density regions, including the valley between modes. As shown in Figure 4.12, 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.

4.8.4. Density–Quantile (DQ) Probability

To obtain a score that is both distribution-aware and comparable across time, this work introduces the 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:

$$\text{DQ}_t = \int_{\{y : p_t(y) \le p_t(x_t)\}} p_t(y)\,\mathrm{d}y. \tag{4.3}$$

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:

$$s_t^{\text{DQ}} = 1 - \text{DQ}_t. \tag{4.4}$$

As shown in Figure 4.12, this score simultaneously suppresses false positives for legitimate operating modes and sharply highlights the injected anomaly located between the modes.

4.8.5. Density–Quantile Severity Scaling

While $1 - \text{DQ}_t$ provides a normalized anomaly score, it does not reflect the relative improbability of extreme events. For example, the difference between $\text{DQ} = 0.99$ and

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:

$$\text{Severity}_t = \min\left(1,\ \frac{p_{\min}}{\text{DQ}_t}\right), \tag{4.5}$$

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 4.12 for $p_{\min} = 10^{-2}$ and $p_{\min} = 10^{-4}$.

4.8.6. Summary

DQ scoring combines likelihood sensitivity with the normalization and time-comparability of quantile-based methods. Unlike residuals, it remains valid under multimodal predictive distributions; unlike QBB, it resolves low-density valleys within the support; and unlike NLL, it yields bounded, comparable scores suitable for persistence tracking, severity ranking, and downstream financial impact estimation.

Therefore, DQ-based scoring is deemed the most suitable anomaly quantification mechanism for MD predictive distributions in this work, directly supporting Objectives **O2** and **O4**, and enabling downstream impact estimation (Objective **O6**).

4.9. Methodological Scope

This chapter derives the core methodological contributions of the thesis: contextual, distribution-aware anomaly detection and scoring under multimodality and non-stationarity (Objectives **O1**– **O4**), conservative signed financial impact quantification (Objective **O6**), and hierarchical localization via asset ontologies (Objective **O5**).

System-level objectives are specified here as design constraints, but are operationalized and validated in subsequent chapters: portfolio-scale execution and tenant isolation in the implementation (Objectives **O7** and **O8**), and robustness to telemetry artifacts in both the ingestion pipeline and benchmarking (Objective **O9**).

5

Benchmarking

This chapter operationalizes the benchmark design principles derived from Chapter 3. In particular, it aims to address the evaluation pathologies highlighted by Liu and Paparrizos [LP24] (Section 3.2) while implementing the domain-specific requirements summarized in Section **??**.

Concretely, the benchmark is constructed to (i) provide clean baseline periods that models can train on or condition on, (ii) represent multivariate contextual point anomalies in the sense of covariate-conditioned detection on consumption targets, and (iii) test seasonal translation by evaluating models under controlled seasonal distance between baseline and evaluation windows.

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 BOPTEST[Blu+21]. 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[IBP25], 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.

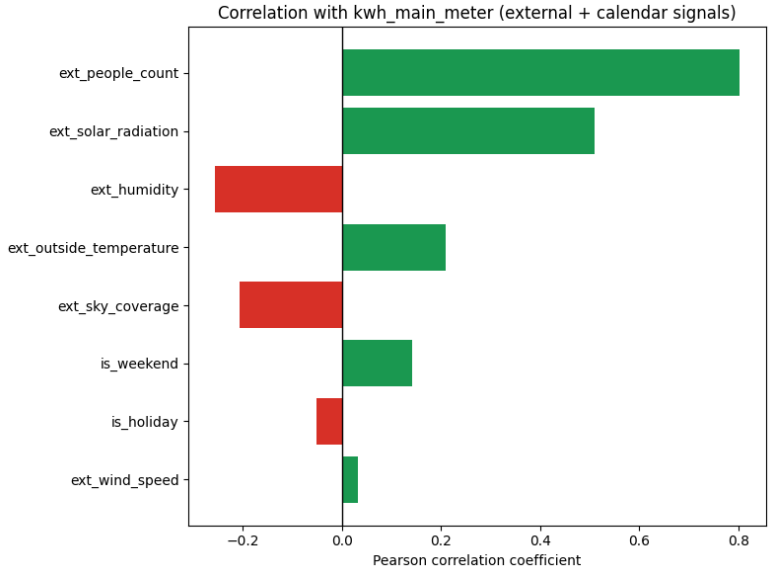


Figure 5.1: Empirical Pearson correlation of contextual and external drivers with the main electricity meter over the created dataset from BOPTEST.

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 (see Section 3.2.1).

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.

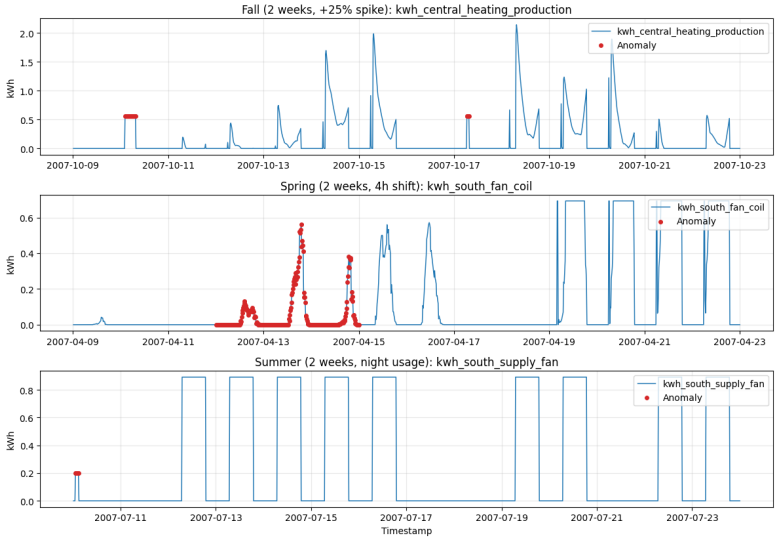


Figure 5.2: Representative sub-meter excerpts from the curated benchmark slices (fall spike, spring pattern shift, summer off-hours). Red markers indicate anomaly windows where the targeted device deviates from its baseline regime.

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. MDNs 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 trained on season-specific clean baseline windows and are evaluated on held-out segments drawn from the same baseline regime, into which synthetic anomalies are injected. Anomaly detection therefore operates relative to a consistent normative reference distribution.

TSFM models, in contrast, cannot be conditioned on the evaluation window itself. They must infer normative behavior exclusively from preceding historical context, which may belong to a different seasonal or operational regime. As a result, detection is performed relative to a shifted baseline distribution rather than the true local operating regime of the evaluation window. This induces a structural evaluation asymmetry that becomes most pronounced in short-window and seasonal translation experiments.

The season-matched three-month baseline configuration provides the closest approximation of comparable operating conditions. In this configuration, Chronos-2 can condition on the complete baseline history within its maximum context length, thereby minimizing 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 TSAD.

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.

Figure 5.3 summarizes the aggregate comparison across benchmark slices, while Figure 5.4 isolates the building main meter to highlight the effect of stochastic output modeling under stable training conditions.

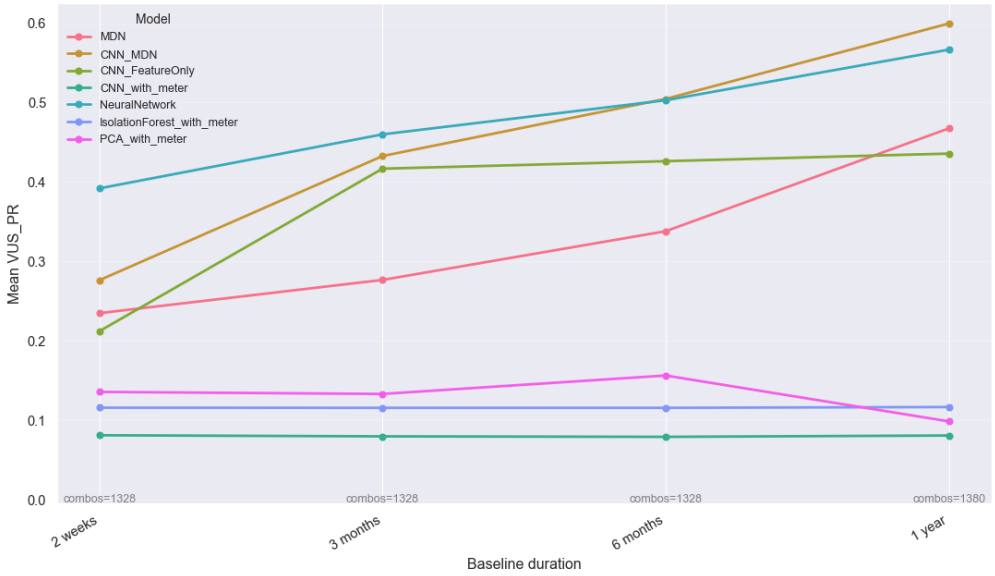


Figure 5.3: Comparison of neural baselines and mixture-density variants across benchmark slices.

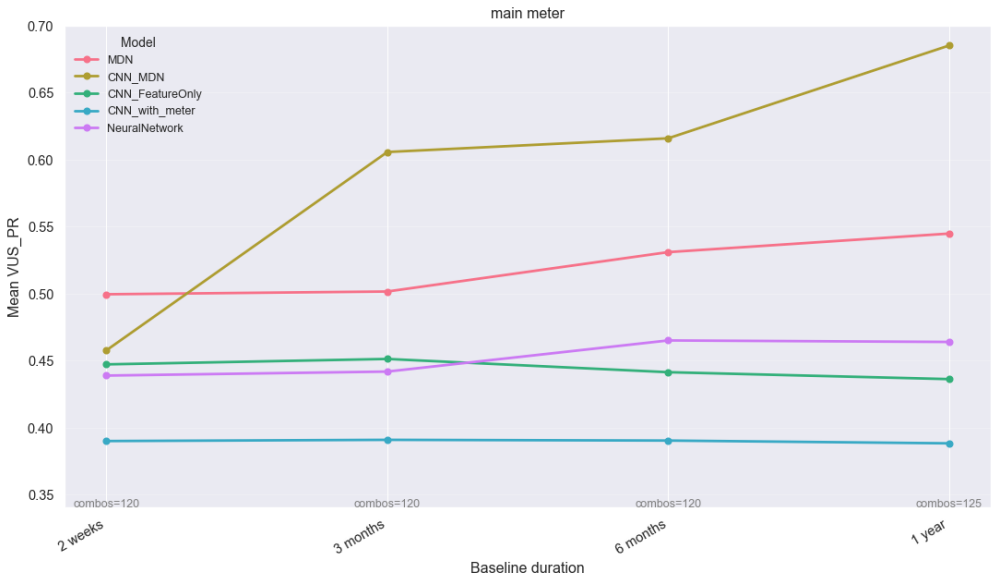


Figure 5.4: Benchmark comparison on the building main meter, highlighting feature-only and stochastic output effects.

5.5.1. Stochastic and Hybrid Architectures

Pure MDNs seem to 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 monotonically with the amount of baseline context available for training in almost all models.

5.5.2. Training Stability and Classical Baselines

MDN layers exhibit sensitivity to initialization and regularization, leading to coverage gaps across meters. principal component analysis (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).

Importantly, the main meter provides a representative example where MDN-based models (including CNN_MDN) converged stably, which is why Figure 5.4 is reported separately. In that setting, MD-based stochastic output modeling yields a substantially stronger VUS-PR than deterministic baselines, suggesting that the weaker aggregate MDN scores in Figure 5.3 are partly a consequence of per-meter instability and reduced coverage rather than an inherent lack of modeling capacity. This limits the strict comparability of aggregated results but still provides evidence for the advantages of probabilistic mixture modeling as discussed in Section 4.8.

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. This configuration is used for the most direct comparison between the foundation model Chronos-2 and trainable baselines, as it minimizes informational asymmetry by aligning the seasonal regime and providing sufficient baseline context. Chronos-2 performs better than most models, with fine-tuning improving it only minimally in off-hours detection. PatchTST[Nie+22] and OmniAnomaly[Su+19] underperform, whereas Autoformer[Wu+21] shows strong performance on spike-like anomalies.

This comparison is visualized in Figure 5.5.

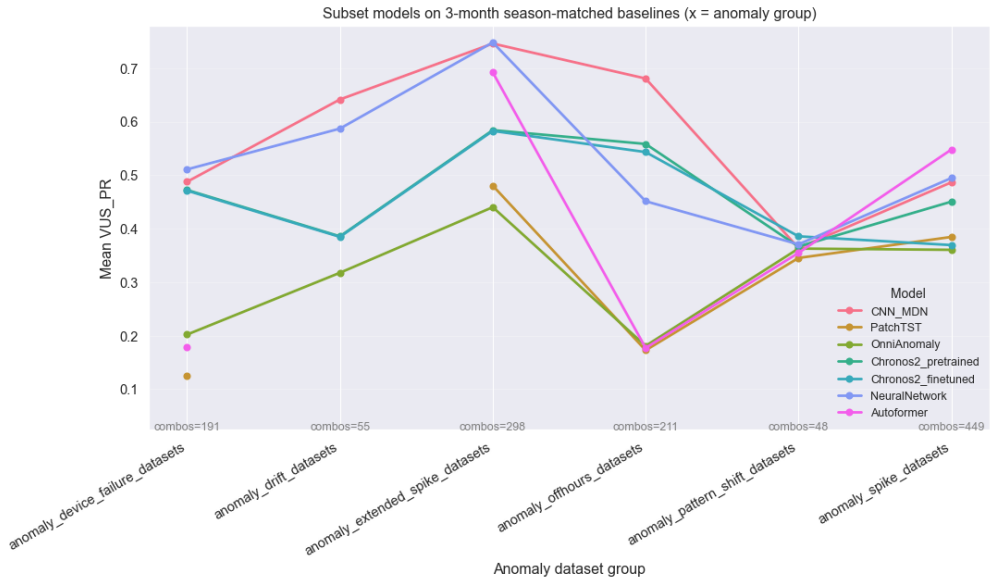


Figure 5.5: Chronos-2 vs. `CNN_MDN` under season-matched three-month baselines.

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 OmniAnomaly, while Autoformer and PatchTST remain comparatively stable across seasonal shifts.

Figure 5.6 summarizes this translation sensitivity.

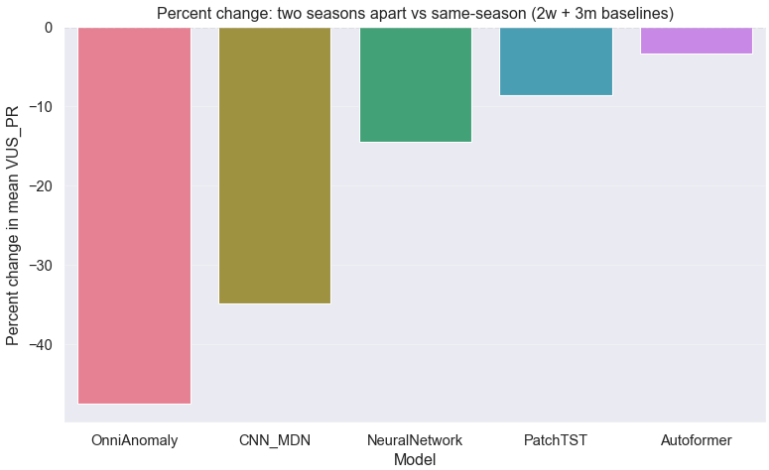


Figure 5.6: Seasonal translation sensitivity of mean VUS-PR versus seasonal distance.

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 because it best satisfies the thesis objectives under portfolio-scale constraints. In particular, the zero-shot, context-conditioned operating mode of TSFMs directly supports **O3** by adapting to baseline drift and abrupt regime changes through recent-context conditioning rather than scheduled per-meter retraining pipelines. This simultaneously strengthens practical scalability (**O7**) and preserves context-conditioned normality (**O1**).

Trainable baselines require explicit per-meter fitting and repeated retraining to track non-stationarity, implying fragile orchestration at scale; persisting anomalies are prevented from contaminating the adaptive reference by applying Strategy B from Section 4.6.4, which operationalizes **O4** in the deployed pipeline.

While Chronos-2 does not expose an explicit multimodal mixture density, it provides calibrated, distribution-free quantile bounds (Section 4.8.2), which satisfy **O2** sufficiently for deployment. These quantile envelopes enable QBB anomaly scoring without parametric assumptions and avoid the instability and coverage gaps observed for mixture-density architectures, while remaining competitive in detection accuracy.

6

Implementation

This chapter summarizes the technical realization of the proposed anomaly detection framework and its integration into the Eliona platform. The implementation is structured to operationalize the methodological and system-level objectives defined in Chapter 4.

6.1. System Integration Overview

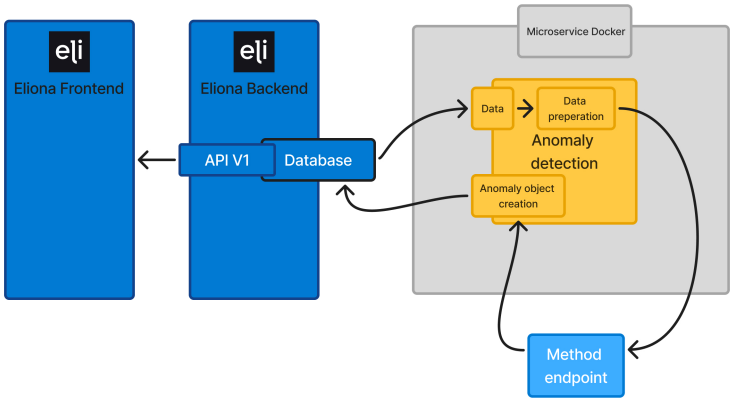


Figure 6.1: High-level deployment of the anomaly detection microservice and Python prediction endpoint within the Eliona and Azure ecosystems.

The framework is implemented as two tightly coupled services:

a Scala-based orchestration microservice responsible for multi-tenant data processing, hierarchical asset handling, and anomaly persistence, and

a Python-based analytics endpoint hosting the Chronos-2 foundation model for probabilistic forecasting.

Both services are containerized and deployed within the same Kubernetes environment as the Eliona backend to ensure low-latency communication and horizontal scalability (**O7**). Logical tenant separation is enforced at all processing stages (**O8**).

6.2. Data Pipeline and Contextual Enrichment

Telemetry is retrieved from the centralized Eliona database and filtered using asset metadata to select economically relevant energy meters. To ensure contextual fidelity (**O1**), each time series is enriched with site-localized weather features prior to inference.

Data integrity is preserved through explicit gap handling and recovery-spike redistribution, preventing non-physical anomaly artifacts caused by telemetry transmission effects (**O9**).

6.3. Probabilistic Inference and Anomaly Quantification

Chronos-2 is queried in batch mode to produce calibrated quantile forecasts for all selected meters. Distribution-free quantile envelopes serve as normative operational bands, fulfilling distributional validity under multimodal and heavy-tailed regimes (**O2**).

To reduce autoregressive reliance, timestamps are evaluated via an h-step forecast (e.g., $h = 10$) using context and known covariates. Detected anomalies are not fed back into the context window but replaced by the predicted expected value, preserving persistence (**O4**).

Anomalies are triggered when observations exceed the extreme quantile envelope. Signed financial impact is computed conservatively using regime-consistent quantile baselines and tenant-specific energy prices, yielding physically defensible monetary loss or savings estimates (**O6**).

6.4. Hierarchical Root Cause Attribution and Action Synthesis

When anomalies are detected on aggregate meters, subordinate assets are evaluated within the dual asset ontology. Signed impacts are aggregated by physical and functional hierarchies, enabling localization of deviations to concrete subsystems (**O5**).

For high-impact events, the resulting diagnostic payload is passed to an LLM-based interpretation layer, which converts structured evidence into human-readable explanations and recommended mitigation actions. The LLM is strictly confined to semantic synthesis and does not influence detection or scoring logic.

6.5. Tenant Configuration and Reproducibility

Tenant-specific parameters control sensitivity thresholds, financial filtering, and energy price calibration. These configurations ensure that detection behavior and alerting thresholds can be aligned with individual operational risk profiles while preserving strict tenant isolation and reproducibility of evaluation results (**O8**, **O7**).

6.6. Frontend Visualization and User Interaction

The frontend operationalizes the backend outputs by exposing persisted anomaly objects as actionable workflows within Eliona's existing navigation (alerts, analytics, asset views). This integration ensures that anomalies can be prioritized by impact (**O6**), interpreted in context (**O1**), and localized to concrete subsystems via hierarchical drill-down (**O5**).

6.6.1. Anomalies List and Operator Validation

Anomalies are surfaced in the alert center via a dedicated *Anomalies* tab. The table provides fast triage using severity, affected asset, timeframe/tags, and signed financial impact, enabling impact-driven prioritization (**O6**). Operators can assign a validation status (e.g., confirmed / false positive) and add comments, supporting long-term trust and iterative refinement under non-stationarity (**O3**).

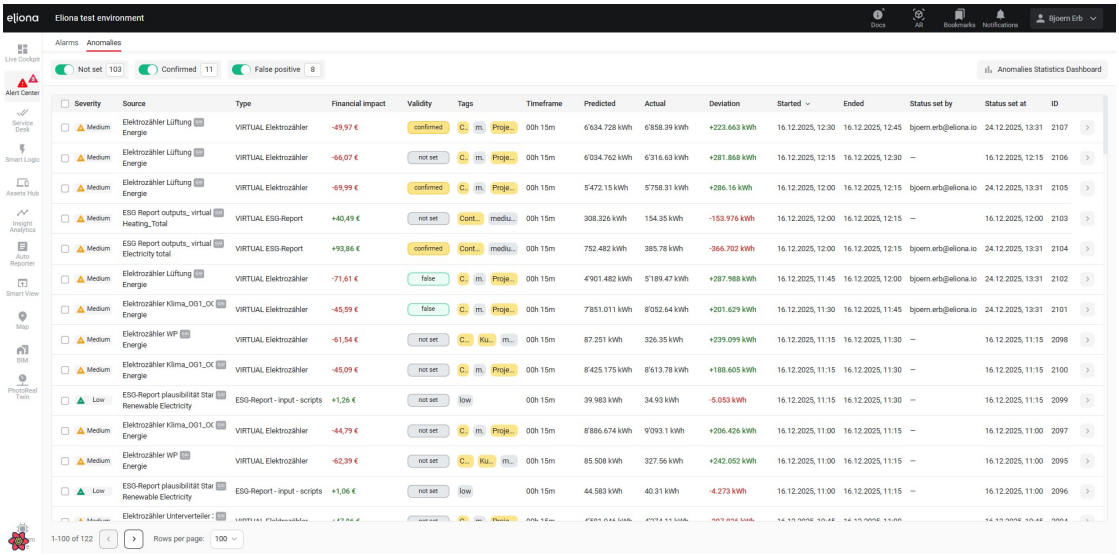


Figure 6.2: Anomalies list interface in the Eliona frontend, showing key diagnostic metrics, financial impact, and filtering options.

6.6.2. Analytics Overlays: Quantile Baselines and Point Highlights

Detection results are embedded into Eliona's analytics framework to avoid siloed tooling. Charts can display anomalous periods and, where resolution matches, point highlights together with the model-derived quantile envelope, making probabilistic baselines visible (**O2**).

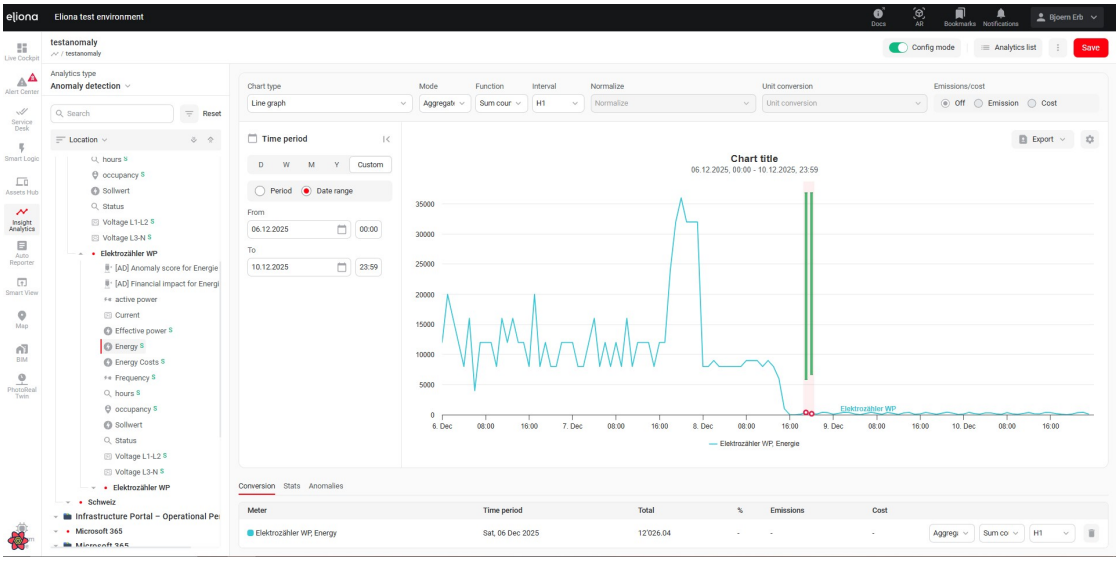


Figure 6.3: Integrated anomaly-detection analytic within the insight analytics framework, highlighting anomalous periods on the consumption charts.

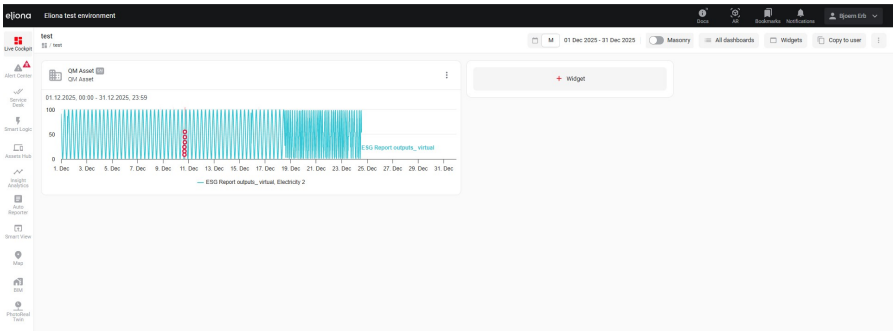


Figure 6.4: Anomaly-detection analytic embedded in a customizable dashboard, combining stochastic overlays with other operational widgets.

Tooltips provide a compact summary at the point of inspection (impact, predicted vs. observed, status, and AI-generated explanation/action), bridging detection outputs and operator decisions.

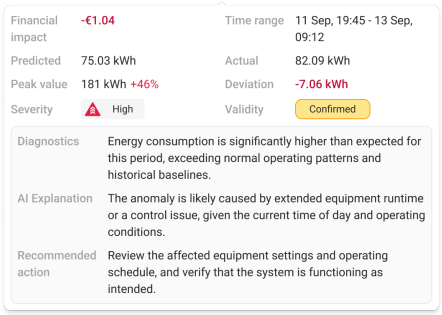


Figure 6.5: Interactive tooltip attached to an anomalous data point, summarizing financial impact, AI explanation, and validation status.

6.6.3. Detail View and Portfolio Reporting

A dedicated detail view consolidates evidence required for remediation: impact and status, an event-centered chart, contextual diagnostics (e.g., weather), and AI-synthesized recommended actions, supporting actionable investigation (**O5**, **O1**).

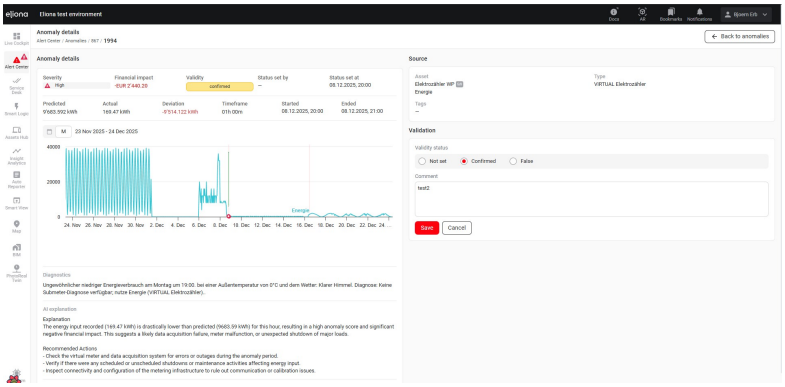


Figure 6.6: Anomaly detail view aggregating financial impact, validation status, analytic chart, diagnostics, and AI-synthesized recommended actions.

For macro-level tracking, an aggregated statistics dashboard summarizes total impact and breakdowns by asset type and time patterns, supporting prioritization and outcome verification (**O6**).

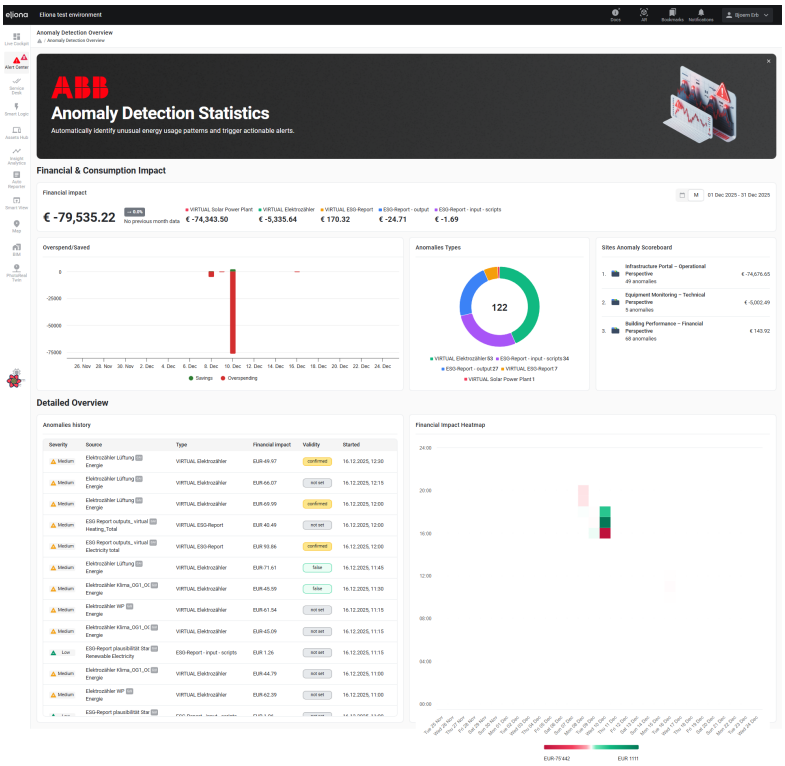


Figure 6.7: Anomaly-statistics dashboard providing macro-level KPIs, categorical breakdowns, temporal trends, and site-level aggregation for executive reporting.

6.6.4. Asset-Level Integration

Finally, anomaly information is embedded into the asset detail view via an *Anomalies* tab, enabling localized investigation directly within the asset context and linking to detailed anomaly views for drill-down (**O5**).

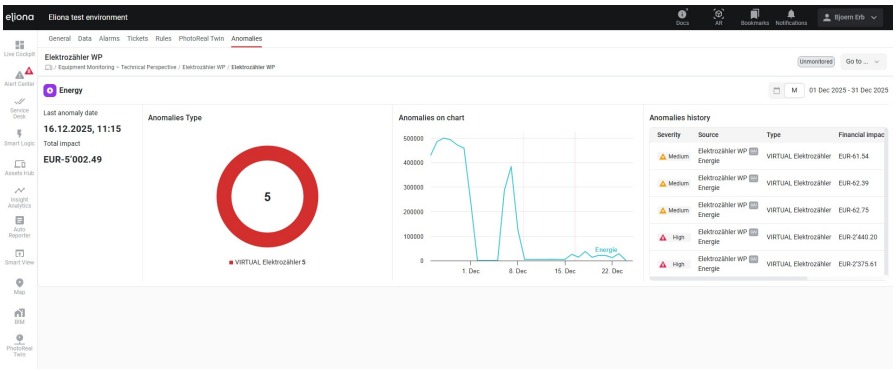


Figure 6.8: Asset-detail view with integrated anomalies tab, showing localized statistics, anomaly-type distribution, and historical events for the selected asset.

6.7. Implementation Summary

The proposed framework is implemented as a containerized Scala orchestration service and a Python forecasting endpoint integrated into the Eliona platform. The design supports portfolio-scale execution with strict tenant isolation (**O7**, **O8**).

Inference inputs are constructed from contextual regime drivers (e.g., localized weather) and validated telemetry, ensuring contextual fidelity and robustness to data artifacts (**O1**, **O9**). Probabilistic quantile forecasts provide distribution-free uncertainty bands for detection under multimodal and heavy-tailed regimes (**O2**), while persistence is preserved through contamination-avoidance and operator validation workflows (**O4**, **O3**).

Detected anomalies are converted into conservative signed monetary impacts and prioritized accordingly (**O6**). RCA localizes deviations via the asset ontology and is surfaced through integrated frontend drill-down, with an LLM restricted to post-hoc explanation (**O5**).

7

Discussion and Future Work

The evaluation of the integrated system demonstrates that combining probabilistic forecasting with hierarchical RCA yields actionable, economically interpretable anomaly management at portfolio scale. At the same time, the transition from a controlled synthetic benchmark to heterogeneous industrial deployments exposes limitations in contextual observability, baseline health assumptions, and the expressiveness of sequential foundation models. This chapter reflects critically on these gaps and outlines concrete directions for strengthening methodological coverage and operational robustness.

7.1. Critical Reflection on System Design

The current RCA implementation localizes anomalies primarily through signed financial attribution across sub-meters (Section 4.5). This is effective for answering *where* excess consumption accumulates and supports **O5** and **O6**, but it provides limited insight into *why* the deviation occurred within the control layer (Foundations Section 2.1 / causal chain discussion in Section 2.1.2). In many facilities, the same consumption pattern can arise from distinct causes (e.g., manual override, schedule misconfiguration, actuator fault). Future iterations should therefore extend diagnostics beyond metering attribution by incorporating *operational state signals* for major consumers (e.g., valve positions, fan speeds, compressor staging, setpoints). This would enable causal consistency checks such as: “is the spike explained by a corresponding change in control demand?” and would strengthen context-aware interpretation (**O1**) while improving the

explainability of high-impact events.

A second design boundary is the *scope of anomaly classes*. The proposed framework is strong at detecting deviations in consumption relative to context-conditioned baselines, but it may not detect *logic conflicts that are consistently present in the baseline*, such as simultaneous heating and cooling, chronic simultaneous reheat, or persistent inefficiencies that have become "normal" operation. This is not a modelling failure but a baseline-definition limitation and directly relates to **O4**.

7.2. Robust Baseline Health Without Manual User Selection

The current production assumption is that historical data preceding activation is sufficiently healthy for establishing a normative reference. In real buildings this is frequently violated: persistent faults may exist for months, and "normal" may already include inefficiency (Foundations Sections 2.1.3 and 2.1.5 on non-stationarity and persistence; Section 4.6 on contamination effects). A manual "gold-standard baseline" UI would shift responsibility to the operator and does not scale across tenants, contradicting the deployment objective **O7**.

A more robust direction is to integrate *classical rule-based energy baselining and detection* as a complementary layer, as discussed in Section 3.1. Rule-based checks (e.g., schedule violations, simultaneous heating/cooling, minimum runtime constraints, holiday/weekend rules) can be used in two ways:

1. **Baseline screening:** automatically tag historical intervals as "likely healthy" vs. "likely faulty" and preferentially sample healthy segments for the model context. This reduces the risk that persistent faults are learned as normal and strengthens **O4** and **O1** without requiring user intervention.

2. **Coverage expansion:** detect fault classes that are not consumption outliers but are operational contradictions (e.g., heating and cooling overlap) even if they were present in the baseline. This provides robustness against unhealthy historical data and complements the probabilistic deviation detector.

Operationally, this hybrid approach would yield a more complete anomaly portfolio: probabilistic scoring for contextual deviations (weather/occupancy adjusted) and deterministic rules for control-logic violations and persistent inefficiencies. The combination directly addresses the failure mode described in Section 4.6 (adaptation and contamination) and improves diagnostic depth beyond pure metering attribution.

7.3. Context and Modality Expansion

The current implementation focuses on electricity meters enriched with meteorological covariates. A natural extension is to generalize the same pipeline to additional utilities (e.g., gas, water, thermal energy, and emissions-related signals), thereby broadening applicability while preserving the same methodological objectives of context-conditioned baselines and signed impact quantification (**O1**, **O6**).

7.4. Future Architecture: Universal Energy Feature Forecaster

Chronos-2 was selected for deployment due to its operational scalability and robust quantile outputs (Section 5.5.5), but the implementation also exposes structural limitations of sequential time-series foundation models:

- **Sequence constraint:** inference requires temporally contiguous context immediately preceding the forecast timestamp, limiting the ability to use non-contiguous "healthy" reference segments (Section 7.2).
- **Autoregressive sensitivity:** sequential dependence can amplify short-term autocorrelation and induce adaptation effects (Section 4.6), requiring mitigation logic in production.
- **Limited distribution access:** forecasts are primarily consumed as quantile bounds, which enables calibrated envelopes but does not expose an explicit multimodal density comparable to MDNs (Sections 4.7.3 and 4.8).

A forward-looking direction is a specialized foundation model designed explicitly as a *universal energy feature forecaster*. Instead of next-step sequence prediction, the model would perform conditional regression from *context features* to a target distribution:

$$p(y \mid \mathbf{x}, \mathcal{B}),$$

where $\mathbf{x}$ are covariates (weather, calendar, occupancy proxies, operational state) and $\mathcal{B}$ is an in-context baseline set of feature–target examples. Critically, $\mathcal{B}$ would not need to be temporally contiguous or adjacent to the forecast time; it could be composed from automatically screened healthy segments across the year (Section 7.2). This directly targets **O1**, **O3**, and **O4** by decoupling the normative reference from immediate sequential inputs.

7.4.1. Mixture-Distribution Output and Distribution-Aware Scoring

To recover the expressiveness of mixture modelling without per-meter training, the model should output an explicit multimodal density (e.g., token-mixture or continuous mixture parameters). This would enable likelihood-based and DQ-based anomaly scoring (Section 4.8) with time-comparable severity scaling, while keeping the zero-shot generalization benefits emphasized in Section 5.5.5. In effect, this would combine the portfolio robustness of TSFMs with the distributional fidelity of MDNs, strengthening **O2** without the instability and coverage gaps observed in trainable mixture architectures.

7.4.2. Decision Support and IPMVP-Style Verification

A feature-conditional, baseline-set forecaster also enables systematic baseline comparison without retraining: the same operating point (same covariates $\mathbf{x}$) can be evaluated under multiple baseline contexts $\mathcal{B}_1, \mathcal{B}_2$ to quantify the expected change attributable to interventions. This naturally aligns with Measurement and Verification practices such as the IPMVP: predicted counterfactual consumption under a pre-intervention baseline can be compared against a post-intervention baseline to estimate savings under matched conditions. This extends the system beyond fault detection into continuous efficiency verification, directly reinforcing the economic motivation of Chapter 1 and strengthening **O6** at portfolio scale. Figure 7.1 summarizes the IPMVP comparison workflow that underpins these verification steps [Eff25].

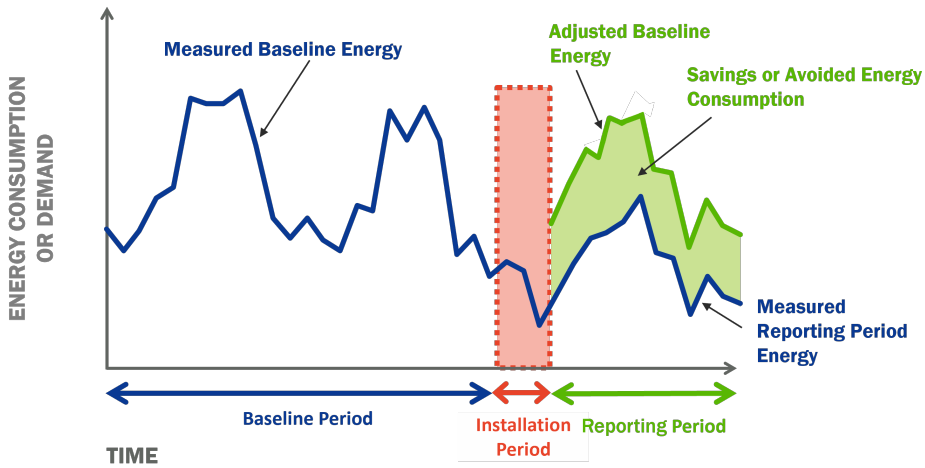


Figure 7.1: IPMVP verification workflow across baseline and reporting periods, adapted from Efficiency Valuation Organization [Eff25].

8

Conclusion

This thesis investigated how portfolio-scale building-energy anomaly detection can be made both operationally deployable and economically actionable. Departing from threshold-based monitoring and deterministic residual detectors (Section 4.6), the problem was formulated as MCAD under multimodality and non-stationarity (Foundations Sections 2.2.1 and 2.1.5), with the explicit requirement that detected deviations be (i) probabilistically well-defined, (ii) monetarily quantifiable, and (iii) diagnostically localizable within building hierarchies (Sections 4.8, 4.4, and 4.5).

Methodologically, the work showed why sequential autoregressive detectors are structurally brittle for sustained building faults: anomalous values contaminate sliding windows, error propagates, and long anomalies can be adapted away (Section 4.6). These failure modes motivate context-conditioned baselines and distribution-aware scoring that remain meaningful across regime changes (Sections 4.6.4 and 4.8). To support scientific comparison under building-relevant anomaly semantics, a domain-specific benchmark was constructed using BOPTEST, providing labeled, contextual fault scenarios with seasonal segmentation and realistic anomaly prevalence (Sections 5.1 and 5.3).

Empirically, the benchmark results reinforced a central design claim: in building energy telemetry, contextual drivers can be more reliable than autoregressive history for establishing normative behavior (Section 5.5). Classical CNN baselines with autoregressive windows performed poorly despite their strong performance in generic multivariate benchmarks, whereas a feature-only CNN relying on exogenous context achieved markedly higher detection scores (Section 5.5). Across up to 16,979 runs per model, the hybrid `CNN_MDN` achieved the highest mean VUS-PR under stable long-horizon training (Section 5.5.1), while mixture-density components exhibited non-trivial initialization and coverage sensitivity across meters (Section 5.5.2), highlighting the

practical importance of robustness under limited tuning budgets.

On the deployment axis, the thesis delivered a production-grade implementation inside the Eliona platform (Chapter 6): a containerized Scala orchestration service and a Python forecasting endpoint integrated in a multi-tenant environment with strict tenant scoping, telemetry integrity handling, conservative signed impact quantification, and hierarchy-aware root cause attribution (Sections 4.4 and 4.5). The diagnostic outputs are surfaced through role-aware workflows (alert triage, analytic overlays, drill-down detail views) and are complemented by an LLM layer restricted to post-hoc semantic synthesis, preserving traceability by keeping detection and scoring deterministic (Section 4.5).

Taken together, the results support three main conclusions. First, building-energy anomaly detection should be treated as a contextual and distributional modelling problem rather than a point-residual thresholding task, because multimodal regimes and non-stationarity dominate real operations (Foundations Section 2.1 and Section 4.8). Second, portfolio-scale value requires economic interpretability and localization (Sections 4.4 and 4.5): without signed impact and hierarchical attribution, detection signals remain difficult to operationalize in maintenance and optimization workflows. Third, evaluation must reflect persistence and contextuality (Sections 4.6 and 3.2); otherwise, reported improvements risk being benchmark artifacts rather than deployable capability.

Finally, the work clarifies where future progress is most impactful. Hybridizing the probabilistic detector with complementary rule-based checks, as established in classical energy monitoring, is a direct path to stronger baseline-health robustness and broader fault-class coverage (Section 3.1 and Section 7.2). Beyond this, a compelling research direction is a universal energy feature forecaster that removes the sequence-contiguity constraint of current time-series foundation models and exposes explicit multimodal densities for distribution-aware anomaly scoring—while enabling standardized baseline comparison aligned with measurement and verification practice (e.g., IPMVP) [Eff25] (Section 7.4).

A

Additional Figures

A.1. Training History Across All Meters

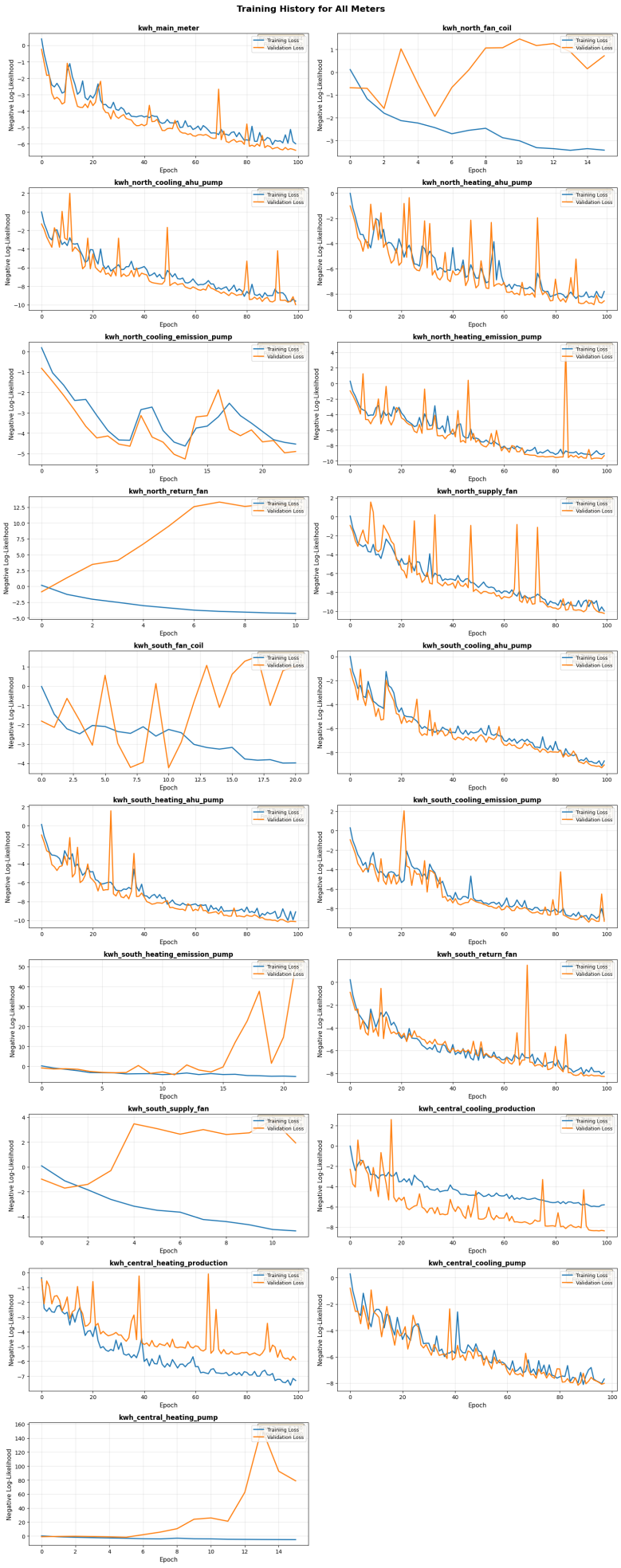


Figure A.1: Training histories aggregated across all meters. This figure is provided for completeness and

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.

[Blu+21] David Blum et al. “Building optimization testing framework (BOPTEST) for simulation-based benchmarking of control strategies in buildings”. In: *Journal of Building Performance Simulation* 14.5 (2021), pp. 586–610. DOI: 10.1080/19401493.2021.1986574. eprint: https://doi.org/10.1080/19401493.2021.1986574. URL: https://doi.org/10.1080/19401493.2021.1986574.

[Wu+21] Haixu Wu et al. “Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting”. In: *Advances in neural information processing systems* 34 (2021), pp. 22419–22430.

[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.

[Nie+22] Yuqi Nie et al. "A time series is worth 64 words: Long-term forecasting with transformers. arXiv 2022". In: *arXiv preprint arXiv:2211.14730* (2022).

[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.

[BPP23] Paul Boniol, John Paparrizos, and Themis Palpanas. *New Trends in Time Series Anomaly Detection: EDBT Tutorial*. Tutorial slides. 2023. URL: https://boniolp.github.io/assets/pdfs/EDBT_tutorial.pdf (visited on 12/28/2025).

[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.

[Eff25] Efficiency Valuation Organization. *International Performance Measurement and Verification Protocol (IPMVP)*. Online protocol. 2025. URL: https://evo-world.org/en/products-services-mainmenu-en/protocols/ipmvp (visited on 12/29/2025).

[Eli25a] Eliona. *Eliona Smart Building Platform*. Website. 2025. URL: https://www.eliona.io/ (visited on 12/27/2025).

[Eli25b] Eliona Smart Building 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).

[Eli25c] Eliona Smart Building Platform. *Geräte mit Eliona verbinden*. Online documentation. 2025. URL: https://doc.eliona.io/collection/connectivity-as-a-service/gerate-mit-eliona-verbinden (visited on 12/28/2025).

[Eli25d] Eliona Smart Building Platform. *Introduction to Ontologies*. Online documentation. 2025. URL: https://doc.eliona.io/collection/eliona-english/academy/introduction-to-ontologies (visited on 12/20/2025).

[Eli25e] Eliona Smart Building 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).

[Eli25f] Eliona Smart Building Platform. *Rules*. Online documentation. 2025. URL: https://doc.eliona.io/collection/eliona-english/documentation/alarms-rules-and-escalations/rules (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.

[IBP25] IBPSA Project 1 BOPTEST. *BOPTEST Test Case Documentation: Multizone Office Complex Air*. Online documentation. 2025. URL: https://ibpsa.github.io/project1-boptest/docs-testcases/multizone_office_complex_air/index.html (visited on 12/28/2025).

[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.