Hochschule Konstanz
Department of Computer Science

Submitted by
Samuel Tim
Student Number 307636

samuel.tim200yahoo.de

Bachelor Thesis

Energy Anomaly Detection with Machine Learning

Konstanz, 31st December 2025

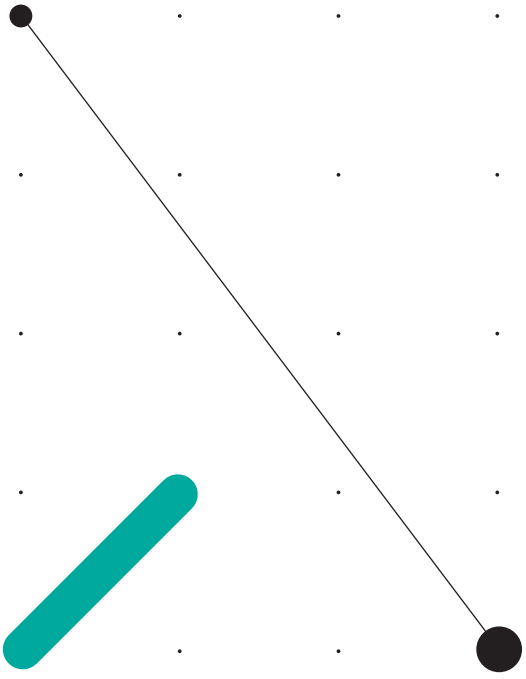
Hochschule Konstanz
Department of Computer Science

Bachelor Thesis

Energy Anomaly Detection with Machine Learning

by

Samuel Tim



in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Science
in Applied Computer Science

at the Hochschule Konstanz University of Applied Sciences,

Student Number:	307636
Date of Submission:	31st December 2025
Supervisor:	**Prof. Dr. Marko Boger**
Second Examiner:	**Dipl.-Inf. Björn Erb**

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

Abstract

Abstract. . .

Contents

Glossary

anomaly observation or pattern that deviates significantly from a defined notion of normality. vi, 9

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

confusion matrix tabular summary of prediction results that counts true positives, true negatives, false positives, and false negatives. vi, 22

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

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

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

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

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

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

Acronyms

BL base load. 9

CNN convolutional neural network. 26–28

DL deep learning. 28

F1 F1 score. 22, 23

FM foundation model. 28

FN false negative. 22

FP false positive. 22

GAN generative adversarial network. 28

HVAC heating, ventilation and air conditioning. 8, 9, 13, 15, 17

IoT Internet of Things. 28

LSTM long short-term memory network. 28

ML machine learning. 27

OmniAnomaly stochastic recurrent neural network model OmniAnomaly. 26, 28

PA-F1 Point-Adjustment F1 score. 26

RE reconstruction error. 28

RP reconstruction probability. 28

SOTA state of the art. 5

Sub-PCA subspace principal component analysis. 26

TN true negative. 22

TP true positive. 22

TSAD time series anomaly detection. 5, 14, 25, 28

TSB-AD Time Series Benchmark for Anomaly Detection. 26, 27

TSFM time-series foundation model. 28

VAE Variational Autoencoder. 28

VUS-PR Volume Under the Surface–Precision Recall. 26, 27

1

Introduction

1.1. Motivation and Problem Statement

Energy efficiency and resource conservation are paramount challenges in the context of global climate change and rising operational costs. In both industrial and residential settings, energy consumption data is being collected at an unprecedented scale, largely due to the widespread adoption of smart meters and Internet of Things (IoT) devices [**SomeSmartMeterPaper2023**].

While this data holds immense value, it also presents a significant challenge: identifying consumption patterns that deviate from the norm. These *anomalies* can represent critical information, such as:

- Equipment malfunction or failure,
- Inefficient operational processes,
- Data integrity issues from faulty sensors, or
- Potential for energy savings and process optimization.

The manual analysis of these vast, high-velocity time series datasets is impractical. Automated methods are required to detect these anomalies reliably and in near real-time. However, defining a ”normal” energy profile is complex due to factors like seasonality (daily, weekly, yearly cycles), weather dependencies, and stochastic user behavior. This complexity makes simple threshold-based detection methods ineffective, leading to a high rate of false positives or missed detections.

1.2. Research Goal and Objectives

The primary goal of this thesis is to design, implement, and evaluate a robust system for anomaly detection in energy consumption time series data. This work aims to compare different algorithmic approaches to identify the most suitable method for a given energy dataset.

To achieve this goal, the following key objectives are defined:

1. **Literature Review:** To investigate the fundamentals of time series anomaly detection and review the current state-of-the-art (Related Work) specific to energy data.

2. **Data Preprocessing:** To select a suitable energy dataset and develop a preprocessing pipeline to handle missing values, normalize data, and engineer relevant features.

3. **Model Implementation:** To implement and train several detection models, likely including a statistical baseline (e.g., ARIMA) and one or more machine learning models (e.g., Isolation Forest, Autoencoder).

4. **Evaluation:** To define and apply appropriate evaluation metrics (e.g., Precision, Recall, F1-Score) to systematically compare the performance of the implemented models.

1.3. Scope and Delimitations

This thesis focuses on **unsupervised anomaly detection**, as labeled anomaly data is rare and expensive to obtain in real-world energy scenarios. The primary input will be univariate energy consumption time series. While external factors like weather or building occupancy are acknowledged as influential, their integration as exogenous variables is considered outside the primary scope of this work but will be discussed as potential future work.

Furthermore, this work is concerned with the **detection** of anomalies, not their **diagnosis**. The system will flag a data point or sequence as anomalous, but it will not perform a root-cause analysis of the anomaly's origin.

1.4. Structure of the Thesis

This document is organized into six chapters:

Chapter 1 (Introduction) motivates the research topic, defines the core problem, and introduces the primary objectives regarding multivariate context point anomalies (MCPA). It further outlines the functional and non-functional requirements for the anomaly detection system.

Chapter 2 (Foundations) establishes the theoretical background for this work. It characterizes building energy telemetry as a multivariate, multi-mode time series, analyzes the causal chain of energy consumption, and examines temporal and statistical properties such as autocorrelation, seasonality, and mixture distributions. The chapter then formalizes anomaly types and learning paradigms, introduces a taxonomy of detection methods with an emphasis on prediction-based approaches, and defines benchmarking concepts including ground truth, confusion matrices, and evaluation metrics.

Chapter 3 (Related Work) reviews and structures the state of the art in time series anomaly detection with a focus on building-related use cases. It discusses the reliability of existing benchmarks, highlighting systemic data flaws and metric biases and motivating the adoption of VUS-PR. The chapter then compares statistical, deep learning, and foundation-model-based methods (including CNN variants, OmniAnomaly, and time-series foundation models such as Chronos-2), and concludes by identifying open research gaps and synthesizing the objectives addressed in this thesis.

Chapter 4 (The Benchmark: Methodology and Selection) presents the design of the custom benchmark tailored to building energy telemetry. It begins with an initial proof of concept to motivate the chosen experimental setup, then details the generation of synthetic datasets with varying baseline lengths (e.g., two weeks, three months, and one year), the labeling strategy for MCPA and global outliers, and the experimental comparison of candidate models such as CNN-based architectures, OmniAnomaly, and Chronos-2. The chapter reports results using the VUS-PR metric and derives the final model choice based on overall performance and seasonal generalization.

Chapter 5 (Technical Realization and System Integration) describes the deployment of the selected model as a high-performance service and its integration into a microservice-based architecture. It explains the Scala-based orchestration of data

preprocessing and anomaly scoring, the bidirectional coupling with the IoT platform, and the persistence strategy for raw telemetry and detected events. The chapter also introduces the root cause analysis logic for attributing anomalies to specific sensors or subsystems and outlines the frontend implementation for monitoring building energy health.

Chapter 6 (Conclusion and Future Work) summarizes the core findings of the research, evaluates the effectiveness of the benchmark and deployed system, and reflects on current limitations. It then sketches a conceptual design for a next-generation multivariate time-series foundation model to address remaining challenges such as seasonal drift and scalability.

2

Foundations

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

2.1. Characteristics of Building Energy Data

To establish the context for effective anomaly detection, the physical properties and behavioral characteristics of building energy data must first be examined. This section provides an analytical study of the data's composition and the underlying factors that determine its structure.

2.1.1. Building Energy as Multivariate Time Series

Energy monitoring produces data in the form of a multivariate time series. To analyze these signals effectively, the mathematical structure of the data must first be defined.

Time Series: A time series is a sequence of data points recorded at successive, typically equal, time intervals. In building automation, these observations represent the continuous state of the system over time.

Multivariate Nature: Unlike univariate data, which only tracks energy consumption, a multivariate time series captures multiple time-dependent variables simultaneously. In this research, the data includes not only the main meter readings but also influencing factors such as outdoor temperature, humidity, occupancy counts, and control setpoints.

Interdependence: The variables in a multivariate building dataset are not independent. Changes in one variable (e.g., an increase in outdoor temperature) lead to changes in another (e.g., cooling power consumption). Effective anomaly detection must therefore account for these cross-variable dependencies rather than treating each series in isolation.

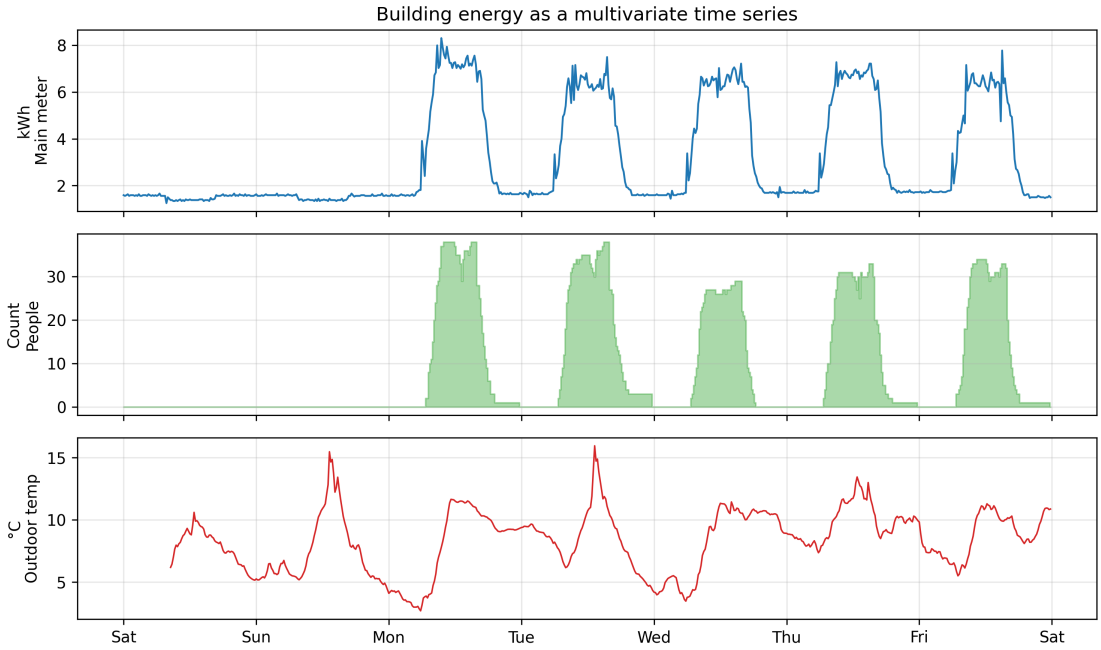


Figure 2.1: Representative multivariate time series showing the main meter load together with occupancy (people count), outdoor temperature, solar radiation, humidity, and selected heating and cooling pumps over one week. The plot illustrates how multiple interdependent variables evolve jointly over time.

2.1.2. The Causal Chain of Energy Consumption

The energy profile of a building is governed by a causal chain that describes the sequential relationship between demand, control, and consumption. Within this frame-

work, every active device—from large-scale industrial machinery to individual lighting units—contributes to the aggregate electrical load.

The overall system operates through interdependent technical layers:

Demand Layer: Environmental factors or occupancy patterns create a requirement for a specific service (e.g., thermal comfort, lighting, or ventilation).

Control Layer: Sensors and controllers interpret this demand and translate it into control signals that regulate the activity of technical systems.

Supply Layer: Mechanical and electrical components activate to fulfill the requirement, resulting in measurable energy use.

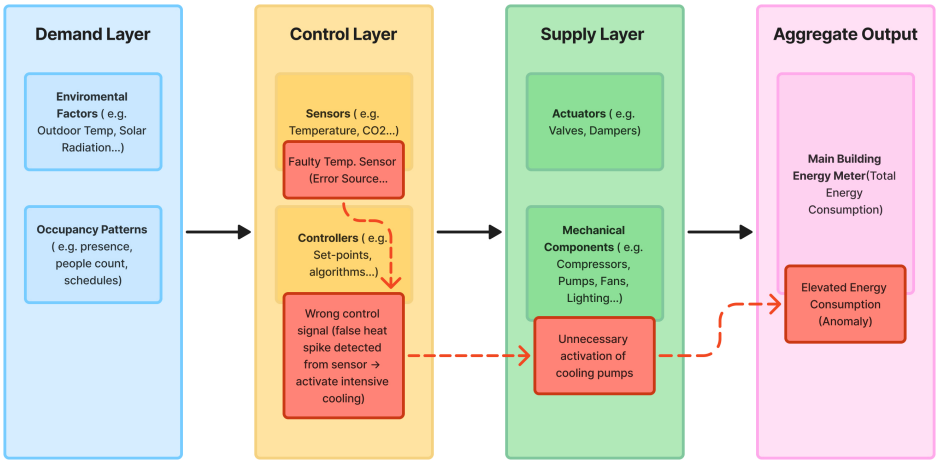


Figure 2.2: Causal chain of building energy consumption from demand over control to supply layer.

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.

Environmental and Technical HVAC Drivers

heating, ventilation and air conditioning (HVAC) systems represent the largest share of energy consumption in most commercial and residential buildings. The operational demand of these systems follows a direct causal sequence from external conditions to mechanical execution:

Meteorological Inputs: The thermal load is directly proportional to the gradient between the setpoint and the outdoor temperature. Solar radiation and humidity further increase the cooling demand on compressors. For instance, high solar gains on a glass facade necessitate intensive cooling even if the outdoor temperature remains moderate.

System Logic and Operational Scheduling: The efficiency of the energy transformation is determined by the control algorithms and predefined temporal patterns. Suboptimal setpoints, such as heating and cooling active simultaneously due to tight deadbands, result in excessive consumption. Scheduling determines the night setback periods; however, if a system heats an empty building at night because of a static schedule, a technically "normal" but inefficient energy load is generated.

Physical System Integrity: The degradation of mechanical components causes a gradual increase in the power required for the same thermal output. Malfunctioning valves or clogged filters increase the static pressure in ventilation ducts, forcing fans to operate at higher speeds. This degradation results in a drift where the energy baseline slowly rises over time.

Internal Loads and Occupancy Dynamics

The internal energy profile is shaped by the presence and behavior of building users, introducing a stochastic element to the data.

Human-Driven Consumption and Thermal Gains: Occupants influence the load directly via lighting and appliances. Additionally, human metabolism and the operation of hardware increase the CO_2 concentration and ambient temperature. This triggers higher ventilation rates and cooling demand.

Behavioral Interventions: Manual actions by users can decouple the energy load from the expected environmental drivers. For example, opening a window during the heating season causes an immediate drop in local temperature, triggering

the HVAC system to ramp up. Conversely, shutting blinds reduces solar radiation, which lowers the cooling load on sunny days.

Integrated IT Infrastructure: Buildings with server rooms exhibit a demand driven by computational load. High server activity increases power consumption and heat production, whereas a system outage results in an immediate drop to the base load (base load (BL)).

Structural and Technical Moderators

The physical environment acts as a moderator for the demand generated by the drivers mentioned above.

Building Envelope and Thermal Inertia: High-quality insulation reduces the energy required to compensate for external fluctuations. The thermal mass of the building prevents instantaneous temperature changes, creating a delay between an external heat spike and the resulting increase in cooling energy.

System Interdependencies: Causal relationships exist between different building services. High-intensity lighting generates waste heat, which increases the load on the cooling system. Therefore, an anomaly in the lighting schedule often manifests as a secondary anomaly in the cooling consumption.

Metering and Data Integrity: The accuracy of the digital record depends on the stability of the measurement infrastructure. Technical noise, such as sensor drift or transmission errors, can create digital artifacts—errors that appear as anomalies in the data but have no physical cause. A common example is a communication rebound: if a meter goes offline and later reconnects, it may send all the "missed" energy data in a single massive spike. Data engineering must therefore distinguish between a physical surge in demand and these digital measurement errors to avoid false alarms.

2.1.3. Temporal Autocorrelation and Persistence

The physical continuity of building systems leads to a high degree of Autocorrelation, where a measurement at a specific point in time is strongly dependent on its preceding values. This relationship is a direct consequence of the building's operational state and physical properties.

Thermal Momentum: The high thermal mass of building structures prevents instantaneous temperature changes. Consequently, the energy required for climate con-

trol is physically linked to the previous state of the system, creating a gradual rather than abrupt transition in demand.

Operational Inertia: Technical systems, such as large ventilation fans or industrial boilers, require time to ramp up or shut down. This results in a continuous consumption curve where subsequent data points remain closely related.

State Persistence of High-Load Devices: Large energy consumers, such as industrial chillers or production machinery, typically operate in sustained cycles. Once a device is activated, it remains in an "on" state for a significant duration to fulfill its operational purpose or to minimize mechanical wear from frequent switching. This creates sustained plateaus of energy demand in the time series.

This high degree of autocorrelation serves as both an advantage and a challenge for anomaly detection. Prediction-based models benefit from this structure because the high dependency on previous values makes the short-term behavior of the system highly predictable under normal conditions. However, this same persistence can hinder the detection of slowly developing faults or persistent higher energy consumption.

2.1.4. Periodic Variations and Seasonality

Building energy data is characterized by seasonality, which refers to regular and predictable fluctuations that recur over fixed intervals. These cycles are a direct consequence of the operational and environmental drivers described in the causal chain and are typically categorized into three temporal scales.

Daily Trends: The 24-hour day–night rhythm is the most dominant cycle, reflecting the primary patterns of solar radiation and standard occupancy. For instance, energy consumption typically ramps up at 07:00 as lighting and ventilation systems activate for arriving staff, and then tapers off in the evening during the night setback phase.

Weekly Trends: These cycles distinguish between standard working days and weekends, resulting in distinct load profiles. A typical office building exhibits high consumption from Monday to Friday, followed by a significant drop on Saturday and Sunday when only the base load—such as emergency lighting and server cooling—remains active.

Seasonal Trends: Annual weather changes shift the primary energy demand between heating and cooling over the course of a year. In a temperate climate, the "heating season" in winter creates a peak in gas or electrical heating demand, while the

"cooling season" in summer causes high electricity consumption for air conditioning and chiller units.

The interaction of these trends creates a complex but repetitive fingerprint of a building's operation. Identifying an anomaly often requires recognizing when a measurement breaks one of these established patterns—for example, if a building shows weekday levels of energy consumption on a Sunday, it indicates a scheduling error in the control layer.

2.1.5. Statistical Distribution and Stochastic Noise

The statistical complexity of building energy data originates from the interaction of discrete system states, environmental extremes, and irregular human behavior. These factors create a distribution that deviates significantly from a standard normal model.

Non-Normal and Mixture Distributions: Energy data rarely follows a Gaussian distribution. Instead, it often manifests as a mixture distribution due to the discrete operational states of technical systems and the superposition of heterogeneous operating regimes. Even when examining a single main meter, the empirical histogram of 15-minute energy consumption typically exhibits multiple local maxima and a long tail, rather than a single symmetric bell curve. When multiple subsystems interact, the aggregate data forms a multimodal profile with several hills, making traditional mean-based detection ineffective.

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.

Causal Ambiguity and Stochastic Coincidence: The high number of interdependent factors in the causal chain makes it difficult to distinguish between true causality and mere correlation. Since multiple stochastic events—such as irregular occupancy, specific weather patterns, and manual device activation—occur simultaneously, a "perfect storm" of unlikely but legitimate events can arise. If several mildly improbable things happen at once, the resulting energy spike may be an unlucky coincidence rather than a technical anomaly.

Sparse State Coverage of Weather and Operation: A significant challenge is the incomplete coverage of the multivariate state space. Historical datasets often lack

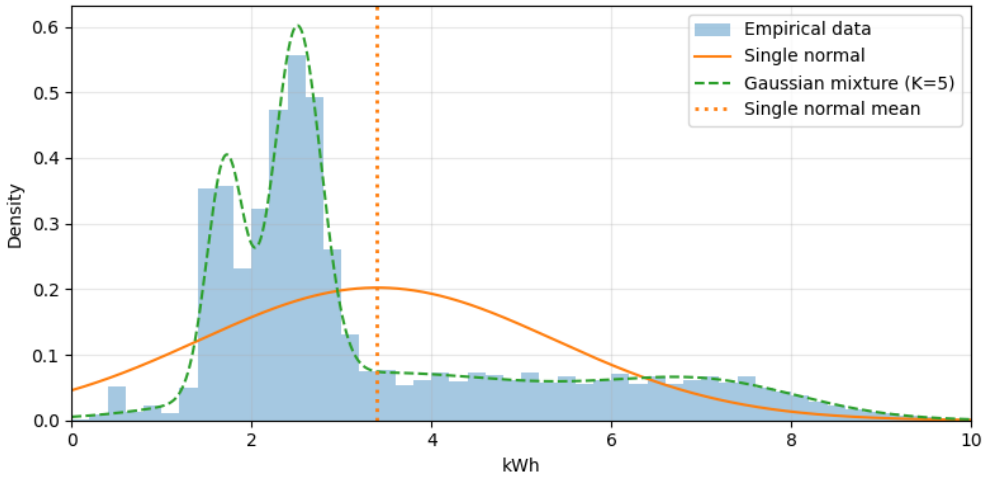


Figure 2.3: Empirical distribution of the building's main meter (15-minute kWh values, histogram) with an overlaid single normal distribution and a Gaussian mixture model with five components, illustrating the mismatch between a unimodal Gaussian model and the multimodal, heavy-tailed structure of real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

observations for extreme meteorological events, such as record-breaking low temperatures or unprecedented heatwaves. Similarly, specific operational corners—such as a full system restart or unique manual overrides—are seldom recorded. When these previously unvisited states are finally encountered, they exist outside the observed distribution and are frequently misidentified as anomalies, despite being physically valid.

Non-Stationarity: Building energy data is frequently non-stationary, meaning its statistical properties, such as mean and variance, evolve over time. Equipment degradation, seasonal transitions, or lasting shifts in occupancy levels push the baseline. Consequently, a model trained on static data may flag normal operational states as anomalous because the underlying distribution has shifted.

2.1.6. Data Acquisition and Semantic Structure

The transition from physical energy consumption to digital analysis follows a multi-staged acquisition process that converts electrical quantities into a structured multivariate time series.

Ontological Data Modeling: To ensure the data is interpretable, the system applies a semantic schema or ontology [Eli25c]. This framework does not merely label individual data points (e.g., "Main Building Meter Consumption") but also encodes the causal chain by defining the physical and logical relationships between me-

ters and devices. By mapping how sub-meters and technical systems, such as HVAC units, contribute to the aggregate load, the ontology provides the contextual intelligence necessary for detection models to localize faults within the technical hierarchy.

Standardized Units: During the transfer from the physical meter to the software, raw readings are converted into standardized units, such as kWh, to ensure consistency across different hardware types.

2.1.7. Data Continuity and Transmission Artifacts

The reliability of the data stream is subject to the stability of the network infrastructure. Interruptions in this chain introduce non-physical distortions that must be distinguished from actual building faults.

Transmission Gaps: Connectivity issues between the building and the software create missing data points. These gaps break the temporal continuity of the record and require correction during the data cleaning stage.

Aggregation Spikes: If a connection is restored after a period of downtime, the system may transmit the entire accumulated energy consumption from the offline period at once. Because the edge gateway recovers buffered data in this manner, the time series can exhibit a virtual spike. These peaks reflect a delay in data reporting rather than a physical surge in energy demand.

2.1.8. Summary of Data Characteristics

The examination of building energy data reveals a multi-layered structure defined by complex physical interdependencies and technical artifacts. These properties establish the operational environment for anomaly detection and determine the requirements for subsequent methodological selection. Because every electrical load is integrated into a causal chain driven by environmental and occupancy factors, effective detection cannot rely on univariate analysis. Instead, models must account for cross-correlations between the primary energy meter and exogenous drivers to resolve causal ambiguity. Furthermore, the presence of mixture distributions and heavy-tailed noise renders traditional mean-based or Gaussian detection methods ineffective. Algorithms must be capable of modeling multimodality and the extreme scarcity of anomalous observations within the multivariate state space.

Temporal characteristics, such as high autocorrelation and recurring seasonal cycles, necessitate models that interpret normality as a time-dependent and context-specific state rather than a static numerical range. Finally, the existence of transmission gaps and virtual spikes requires a clear distinction between physical system malfunctions and digital artifacts. Consequently, data integrity is not an inherent property of the telemetry stream but a condition that must be established through robust preprocessing. These characteristics provide the necessary context for the formal principles and taxonomies of anomaly detection.

2.2. Foundations of Anomaly Detection

Anomaly detection is the process of identifying observations or patterns that deviate significantly from a defined notion of normality. In the specific context of time series data, an anomaly is defined not merely by its numerical value, but by its relationship to the temporal sequence. Unlike static data analysis, where outliers are detected in an independent and identically distributed feature space, TSAD must account for the ordering, dependency, and trend-based characteristics of the signal.

The theoretical definitions and taxonomies established in this section follow the benchmark framework proposed by Paparrizos et al. [Pap+22]. Beyond the presentation of these fundamental concepts, this section systematically classifies the building energy data utilized in this research within the established taxonomies to define the specific requirements for the detection framework.

2.2.1. Dimensionality and Modes of Normality

The complexity of the detection task is governed by the structural type of the time series and the diversity of its underlying operational regimes. Based on the dimensionality and the number of normal states, the data is classified as follows:

Dimensionality: Univariate vs. Multivariate: A time series is univariate if it consists of a single time-dependent variable, such as the aggregate energy consumption of a building. In contrast, multivariate time series capture multiple interdependent variables simultaneously. The dataset utilized in this research is strictly multivariate, as it integrates energy consumption with exogenous drivers like outdoor temperature and occupancy. This multidimensional approach is necessary to resolve causal ambiguity; for example, a high cooling load is only interpretable when compared against a concurrent heatwave or high occupancy count.

Modality: Single-Mode vs. Multi-Mode Normality: A system exhibits single-mode normality if its behavior follows a consistent, unified pattern. However, building energy systems typically operate in multi-mode normality regimes. This means that "normal" behavior shifts depending on the operational context. An intuitive example is the seasonal transition of an HVAC system: a high electrical load during a summer afternoon is a normal response to cooling demand, whereas the same load profile in winter—where heating is primarily gas-driven—would be highly anomalous.

The data in this work is therefore characterized as a multivariate, multi-mode time series. This classification necessitates detection algorithms that can learn complex cross-variable correlations and adapt to shifting baselines without generating false positives during seasonal transitions.

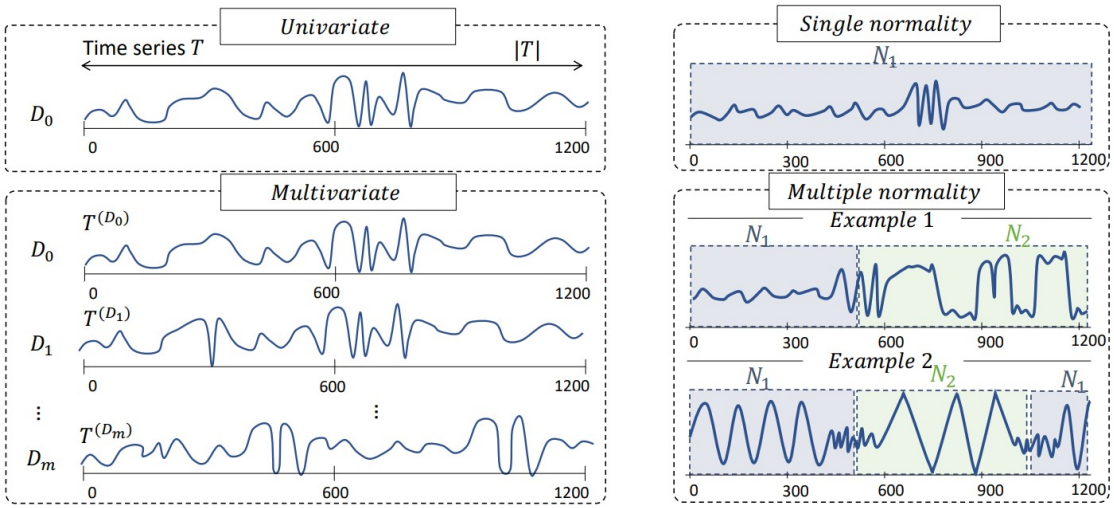


Figure 2.4: Schematic illustration of time series types along two axes—dimensionality (univariate vs. multivariate) and normality regimes (single-mode vs. multi-mode). Adapted from Boniol et al.'s tutorial on new trends in time series anomaly detection [**Boniol2023NewTrends**].

2.2.2. Taxonomy of Anomalies

Anomalies are classified based on their structural characteristics within the time series and their patterns of occurrence. This systematic categorization, as illustrated in Figure 2.5, is essential for understanding the nature of deviations and selecting appropriate detection methodologies.

Structural Classification

The structural classification differentiates between anomalies that appear as individual data points and those that manifest as sequences over time.

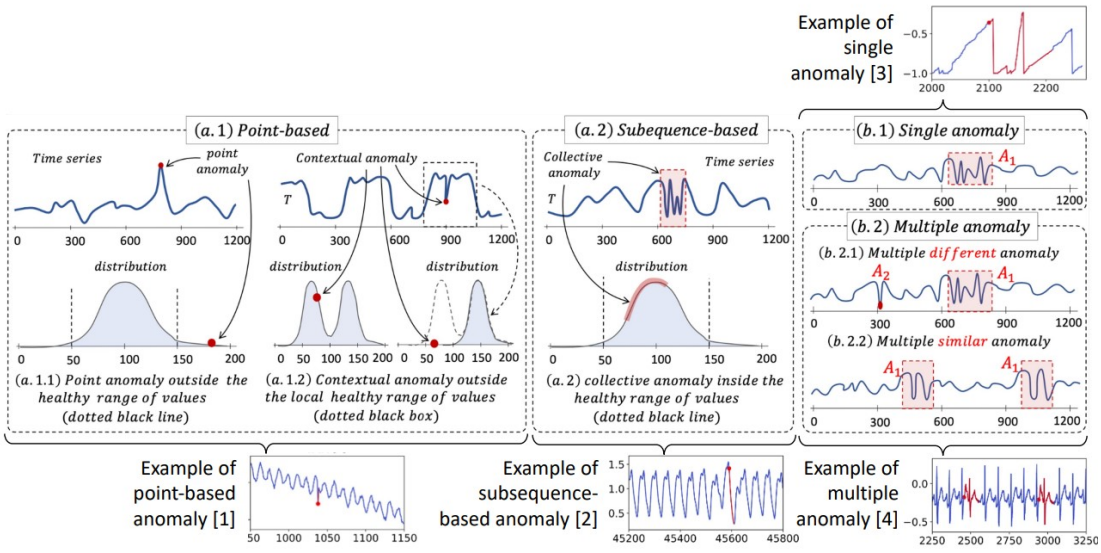


Figure 2.5: Taxonomy of time series anomalies along structural and multiplicity dimensions, distinguishing global and contextual point anomalies, subsequence-based anomalies, and their occurrence as single, multiple different, or multiple similar events. Adapted from Boniol et al.'s tutorial on new trends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

Point-Based Anomalies: Individual observations that deviate from the expected behavior. Multiple such anomalous points may occur consecutively; this alone does not constitute a subsequence-based anomaly but remains a series of point anomalies as long as each timestamp can be evaluated independently.

Global Point Anomalies: A data point is considered a global anomaly if its value falls entirely outside the range of healthy values observed in the historical dataset (see Figure 2.5 a.1.1). In building energy data, such extremes are often caused by sensor malfunctions or virtual spikes from communication rebounds.

Contextual Anomalies: A data point is considered a contextual anomaly if it deviates from the expected behavior within a specific local context, even if its value lies within the global healthy range (see Figure 2.5 a.1.2). The context is typically defined by temporal attributes (e.g., time of day, weekday vs. weekend) or external covariates such as weather and occupancy.

Subsequence-Based Anomalies: Anomalies that arise when a contiguous sequence of data points exhibits abnormal behavior, even if the individual points within that sequence are not necessarily anomalous on their own (see Figure 2.5 a.2). These are often referred to as collective anomalies, where the pattern of the sequence itself is deviant [Pap+22].

Multiplicity and Similarity

Anomalies are further categorized based on their frequency and similarity within the dataset.

Single Anomalies: Isolated anomalous events that occur at a specific point in time or within a single interval (see Figure 2.5 b.1).

Multiple Anomalies: Situations in which several anomalous events occur within the dataset.

Multiple Different Anomalies: The occurrence of various distinct anomaly types (see Figure 2.5 b.2.1), such as a combination of sensor faults and control logic errors.

Multiple Similar Anomalies: The recurrence of the same anomaly pattern over time (see Figure 2.5 b.2.2), often indicative of a persistent fault or systematically inefficient operation.

Classification within the Research Context

The research presented in this work primarily focuses on the detection of contextual point anomalies and addresses the challenges posed by multiple similar anomalies.

Emphasis on Contextual Points. The detection framework prioritizes contextual point anomalies to account for the influence of external drivers. For example, an extraordinarily cold day may result in energy consumption that reaches a historical maximum. While this would be flagged as a global point anomaly, it represents a normal response to the environmental context. Conversely, a light left on overnight might yield a value that appears low on a global scale but is significantly higher than expected for the nighttime context. True anomalies are therefore defined by their deviation from the expected load given the prevailing conditions.

Manifestation of Sequence Deviations. High-frequency subsequence anomalies, such as an HVAC unit switching on and off every minute, are also captured as contextual point anomalies in aggregated energy data. Although the fault originates as a rhythmic sequence, the cumulative effect of frequent starts produces an interval total (e.g., over 15 minutes or one hour) that is significantly higher than typical. Because substantial subsequence deviations manifest as detectable point outliers in the aggregated series, the methodological focus remains on the identification of contextual point anomalies.

Challenge of Recurring Patterns. Systematic faults frequently produce multiple similar anomalies. A central challenge is the risk of normality drift: if an anomaly occurs with high regularity, the detection model may gradually incorporate the deviation into its learned definition of normal behavior. Without safeguards, this adaptation suppresses alerts despite the presence of a persistent inefficiency.

2.3. Methodological Approaches to Anomaly Detection

The detection of anomalies transforms raw time series into actionable intelligence by quantifying deviation and selecting algorithms that match the available supervision signal.

2.3.1. The Anomaly Score

Most anomaly detection methods are capable of generating an anomaly score for each individual data point in the series as an output. This output score is typically normalized to a range between 0 and 1, where higher values represent an increased probability of anomalous behavior.

To convert these continuous scores into actionable alerts, a numerical threshold is applied to the resulting sequence. If a threshold of 0.8 is selected, all data points with a score exceeding this limit are classified as anomalies, while values below the threshold are categorized as normal. As illustrated in Figure 2.6, the anomaly score can be visualized alongside the underlying time series, with the threshold line separating nominal from anomalous observations.

2.3.2. Learning Paradigms in Energy Data

The selection of an anomaly detection methodology is determined by the availability and nature of labeled training data. These strategies are categorized into supervised, semi-supervised, and unsupervised paradigms.

Supervised Learning: Supervised approaches utilize training datasets that contain explicit labels for both normal operating states and specific anomaly examples (see Figure 2.7). In the domain of building operations, this paradigm is rarely applicable because facility managers and engineers seldom provide precise annotations of historical faults.

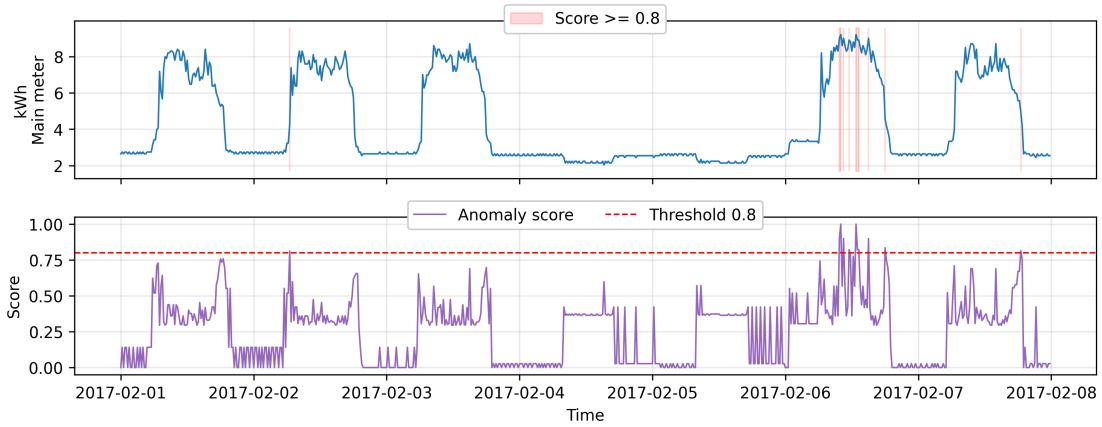


Figure 2.6: Example of an anomaly score $s_i \in [0, 1]$ aligned with the underlying time series. A threshold of 0.8 separates normal points from those flagged as anomalous.

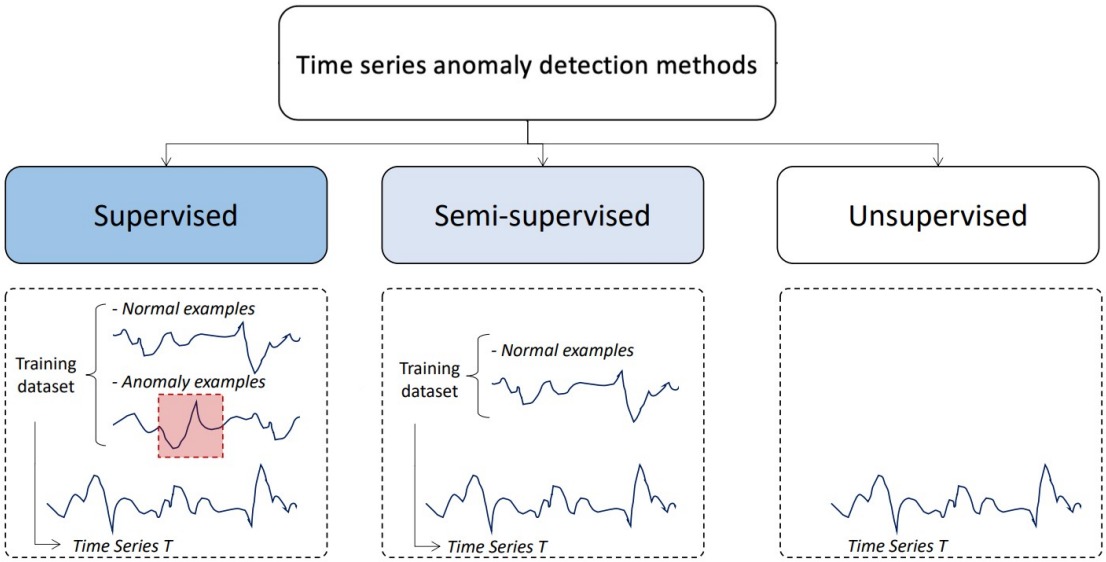


Figure 2.7: Schematic overview of supervised, semi-supervised, and unsupervised learning paradigms for anomaly detection, adapted from Boniol et al.'s tutorial on new trends in time series anomaly detection [**Boniol2023NewTrends**].

Semi-Supervised Learning: Semi-supervised learning relies on a training dataset composed exclusively of healthy or normal examples (see Figure 2.7). This set serves as the baseline for the model to learn the characteristic patterns of a specific building. This approach is frequently employed in building energy management when historical data is agreed upon as a healthy reference or when the assumption is made that the majority of past operations were conducted without significant technical failure.

Unsupervised Learning: Unsupervised learning operates without any prior labels or dedicated training phases based on healthy data (see Figure 2.7). The algorithm identifies anomalies by searching for statistically rare events or structural deviations within the current time series itself. This environment is typically encountered during the initial commissioning of a building when no historical record exists to establish a baseline.

2.3.3. Taxonomy of Detection Methods

Methodological approaches to anomaly detection are categorized into functional families based on their underlying logic. These families include distance-based, density-based, and prediction-based methods, as illustrated in the hierarchical taxonomy in Figure 2.8.

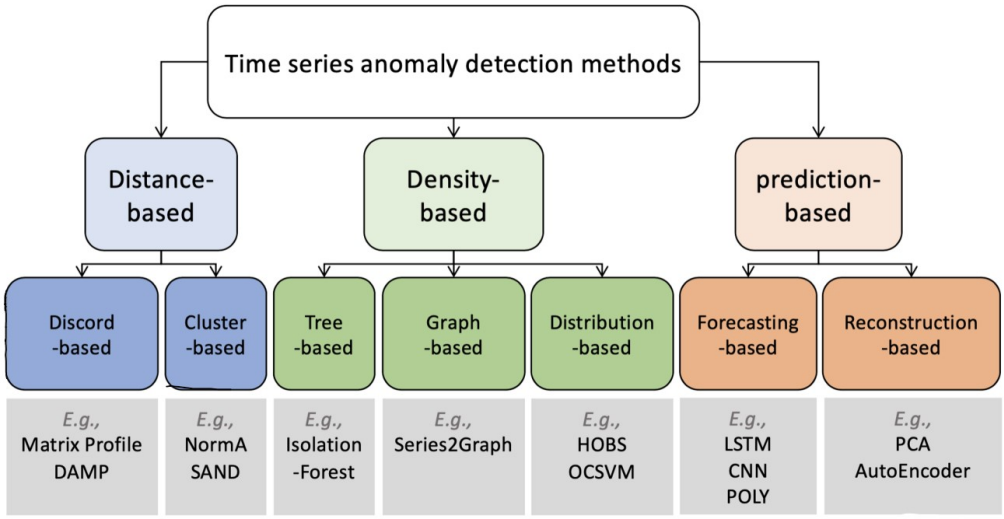


Figure 2.8: Hierarchical taxonomy of anomaly detection methods grouped into distance-based, density-based, and prediction-based families, adapted from Boniol et al.'s tutorial on new trends in time series anomaly detection [**Boniol2023NewTrends**].

Distance-Based Methods: These techniques identify anomalies by measuring the similarity between data sequences. Subsequences that exhibit a high distance from all other patterns in the dataset, often referred to as Discords, are classified as anomalous.

Density-Based Methods: This approach evaluates the local density of data points within a feature space. Observations located in sparse regions, where the number of neighboring points is significantly lower than the average density, are identified as outliers.

Prediction-Based Methods: These algorithms identify deviations by analyzing the difference between a model's generated representation of a signal and the actual observed measurement. This family encompasses Forecasting-based techniques, which typically utilize historical data to predict future values, and Reconstruction-based techniques, which learn to recreate the input data itself.

The research presented in this thesis focuses specifically on prediction-based methodologies. The technical justification for this prioritization, particularly regarding the integration of multivariate environmental drivers and the establishment of energy baselines, is provided in the subsequent chapters.

2.4. Benchmarking Foundations

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

2.4.1. Binary Labeling and Ground Truth

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

2.4.2. The Confusion Matrix: Four Possible Outcomes

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

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

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

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

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

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

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

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

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

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{2.1}$$

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

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

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{2.2}$$

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

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

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.3}$$

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

2.5. Synthesis of Foundations

The examination of building energy data in Section 2.1 and the formal taxonomies of anomaly detection in Section 2.2 reveal a highly specialized operational environment. Because building energy consumption is an aggregate signal driven by a complex causal chain (see Subsection 2.1.2), it is fundamentally characterized as a multivariate time series (see Subsection 2.1.1). This multidimensionality necessitates a focus on contextual point anomalies (see Subsection 2.2.2), as abnormality is defined primarily relative to the prevailing environmental and operational state rather than to a single global range. Furthermore, the temporal aggregation of energy data into 15-minute or hourly intervals (see Subsection 2.1.5) effectively transforms high-frequency rhythmic faults into detectable point-based deviations in the aggregated series.

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

Within this context, semi-supervised and unsupervised learning paradigms (see Subsection 2.3.2) introduce a fundamental technical contradiction when historical data is utilized in the absence of expert-annotated fault histories. If persistent or repetitive anomalies were already present during the training period, the model incorrectly incor-

porates these deviations into its learned definition of normality. This risk of normality drift, particularly in the presence of multiple similar anomalies (see Subsection 2.2.2), leads to suppressed alerts and remains a central constraint for the subsequent methodological selection.

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

3

Related Work

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

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

3.1. Reliability and Benchmarking: The TSB-AD Framework

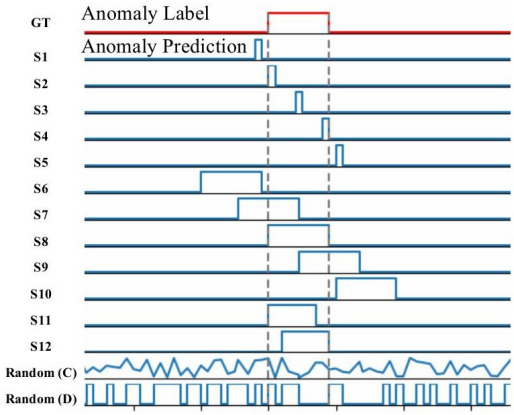
The selection of an appropriate detection methodology is constrained by systemic issues within the existing research landscape. Liu and Paparrizos [LP24] identify these issues as the "elephant in the room," demonstrating that apparent progress in TSAD is often an artifact of flawed evaluation practices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

3.1.1. Systemic Flaws and Metric Reliability

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

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

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



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].

3.1.2. Benchmark Evaluation and Model Hierarchy

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

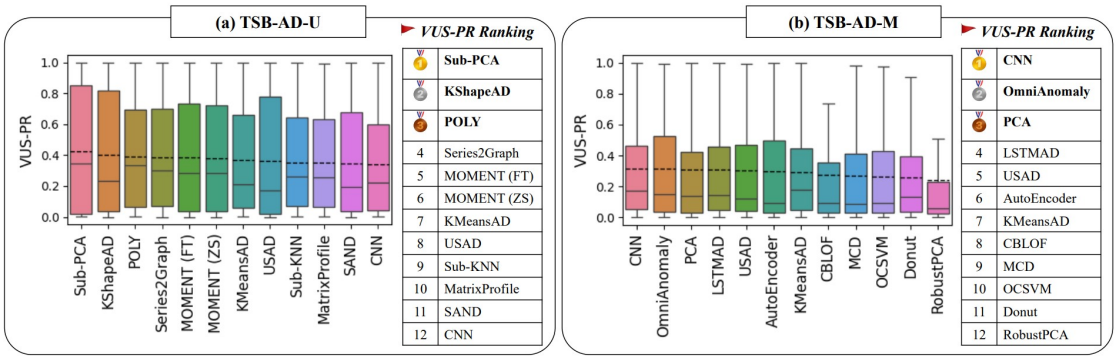


Figure 3.2: Accuracy evaluation of the top 12 methods on (a) univariate (TSB-AD-U) and (b) multivariate (TSB-AD-M) datasets based on the VUS-PR metric. Adapted from Liu and Paparrizos [LP24].

3.1.3. Implications for Multivariate Context Point Anomalies

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

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

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

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

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

3.2.1. Deep Generative Models and the Advantage of Reconstruction

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

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

3.2.2. The Emergence of Time-Series Foundation Models

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

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

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

3.2.3. Synthesis: Prediction-Based Stochastic Modeling

A synthesis of current literature reveals a converging trend toward prediction-based methodologies as the primary solution for multivariate context point anomalies in building energy data (see Section 2.3.3). While reported results remain susceptible to the data-level deficiencies and biased metrics identified in Section 3.1.1, the scientific consensus points toward a transition from deterministic modeling to stochastically adjusted prediction. It was shown through the findings of Azzalini et al. [Azz+25], Su et al. [Su+19], and Ansari et al. [Ans+25] that robust results are achieved when a model outputs a probability distribution rather than a single point value.

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

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

3.3. TODO: Mixture Density Networks for Stochastic Modeling

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

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

3.4. TODO: Root Cause Analysis for Anomaly Detection

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

3.5. Identification of Research Gaps

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

3.5.1. Representation Gap in Multivariate Context Point Anomalies

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

3.5.2. Temporal Baseline and Generalization Gap

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

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

3.5.3. Architectural Gap in Stochastic Mixture Modeling

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

3.5.4. Diagnostic and Economic Functional Gap

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

3.6. Synthesis of Research Objectives

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

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

4

Methodology

4.1. System Context: The Eliona IoT Platform

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

4.1.1. Modular System Architecture

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

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

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

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

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

4.1.2. Asset Modeling and Hierarchical Ontology

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

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

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

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

4.2. System Requirements Specification

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

4.2.1. Functional Requirements

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

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

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

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

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

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

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

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

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

4.2.2. Operational and Data Integrity Requirements

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

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

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

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

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

4.3. Proposed High-Level Architecture

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

4.3.1. Component Interaction and Data Flow

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

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

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

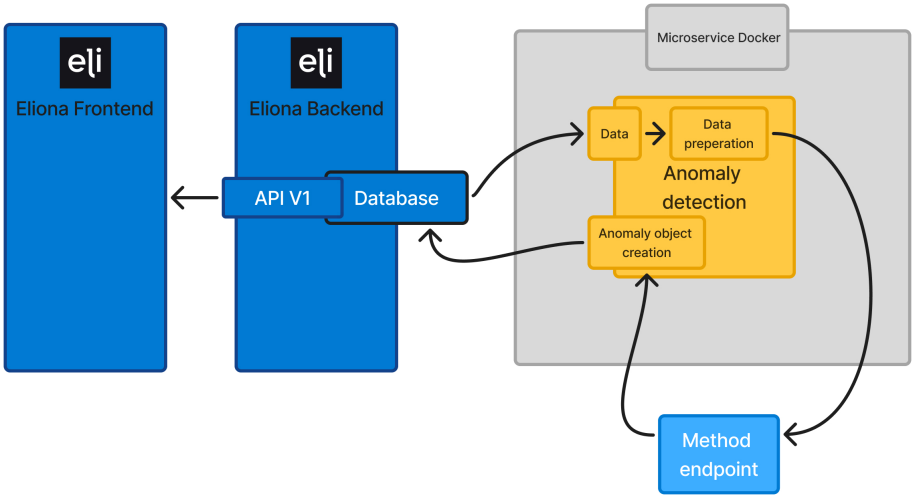


Figure 4.1: High-level architecture of the integrated anomaly detection system, illustrating the interaction between the Eliona Core, the anomaly detection microservice, and the analytics endpoint.

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

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

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

4.3.2. Integration into the Azure Ecosystem

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

4.4. Proof of Concept (PoC): Stochastic Feasibility

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

4.4.1. MDN-Based Stochastic Prediction

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

4.4.2. Diagnostic Integration: SHAP and LLM

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

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

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

4.5. Critique of Sequential Forecasting for Anomaly Detection

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

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

4.5.1. Synthetic Experimental Setup

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

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

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

4.5.2. Failure Mode 1: Error Propagation and Instability

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

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

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

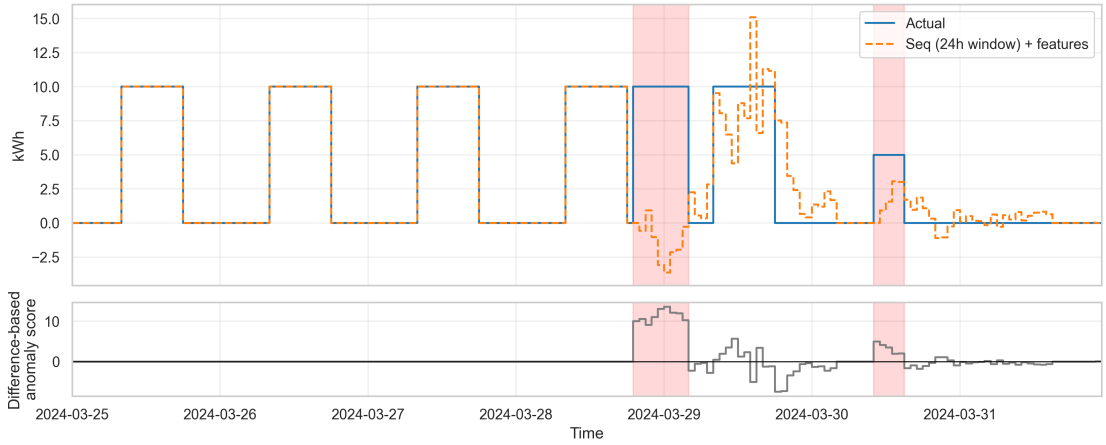


Figure 4.2: Prediction behavior of a model using a 24 h historical window plus time features. The top panel shows actual vs. predicted values; the bottom panel shows the difference-based anomaly score. Note the erratic predictions even after the anomaly ends as the unseen data propagates through the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

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

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

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

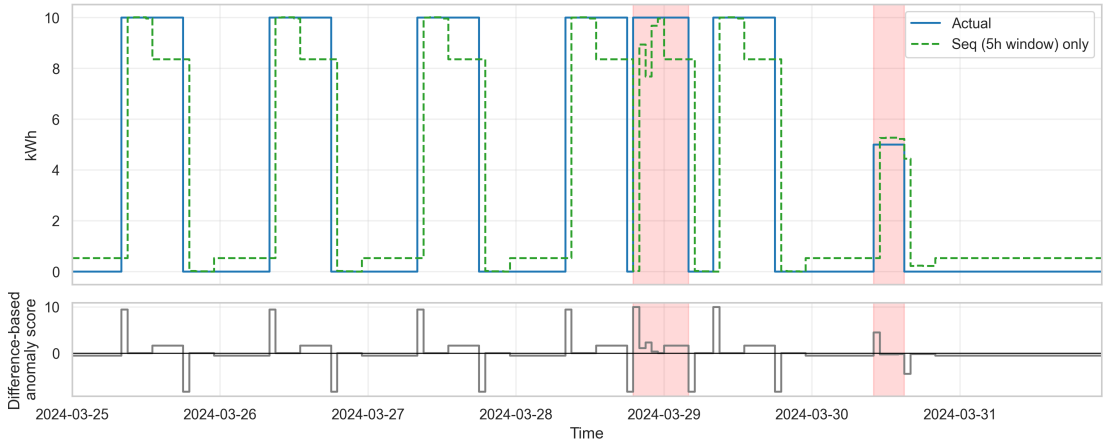


Figure 4.3: Prediction behavior of a model using only a short (5-step) historical window. The model correctly identifies the onset of anomalies but rapidly adapts to the new level, causing the anomaly score to drop back to near zero while the anomaly is still ongoing.

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

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

Implications for Evaluation Metrics and Financial Impact

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

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

4.5.4. Mitigation Strategies

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

Strategy A: Contextual Feature-Only Modeling

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

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

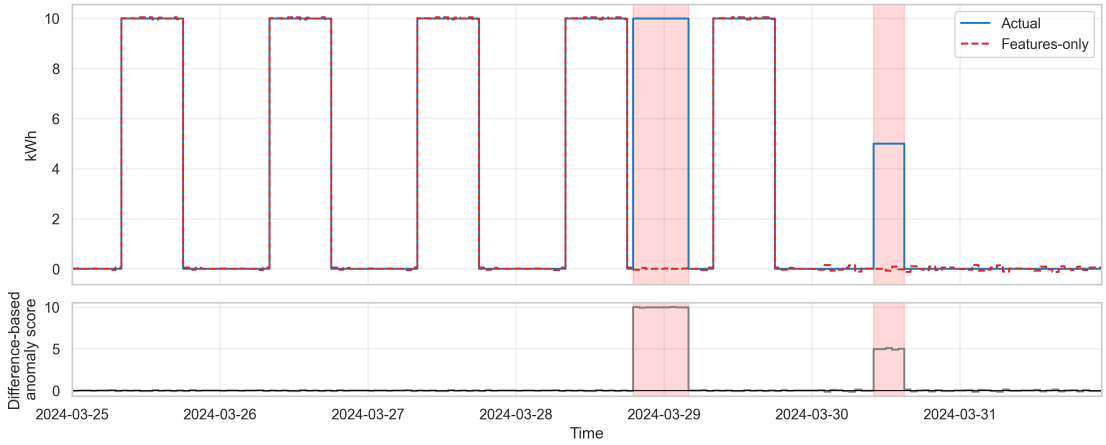


Figure 4.4: Behavior of a features-only model (no historical consumption input). The prediction relies solely on context (time/weekend), resulting in a stable baseline and accurate detection of sustained anomalies without adaptation.

Strategy B: Inference-Time Input Imputation

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

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

4.6. Statistical Limitations of Point and Gaussian Predictions

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

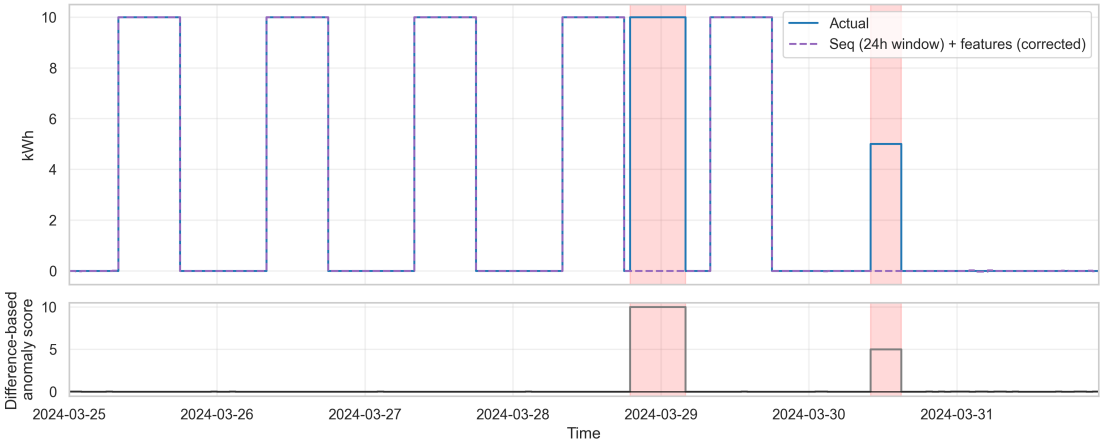


Figure 4.5: The same 24-hour window model from Figure 4.2, but applied with inference-time imputation. When an anomaly is detected, the predicted value replaces the actual value in the sliding window for future steps. This prevents error propagation and maintains a high anomaly score throughout the event.

scoring behavior diverges drastically.

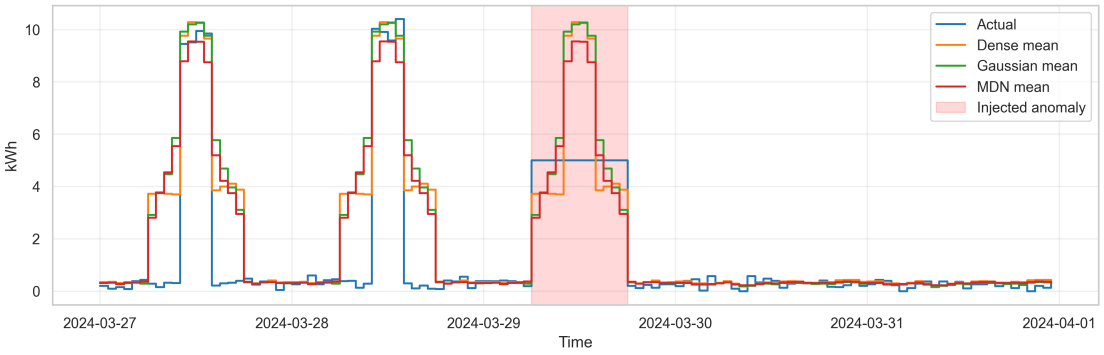


Figure 4.6: Predicted means over the Variable Shift horizon. Dense, Gaussian, and MDN models track the two regimes, masking the scoring deficiencies discussed in Sections 4.6.1–4.6.3.

4.6.1. The Failure of Mean Squared Error Minimization

Dense regressors trained with Mean Squared Error (MSE) converge toward the global average of both regimes. In bimodal settings this leads to systematic bias: the model predicts approximately $5\,\text{kWh}$ 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.

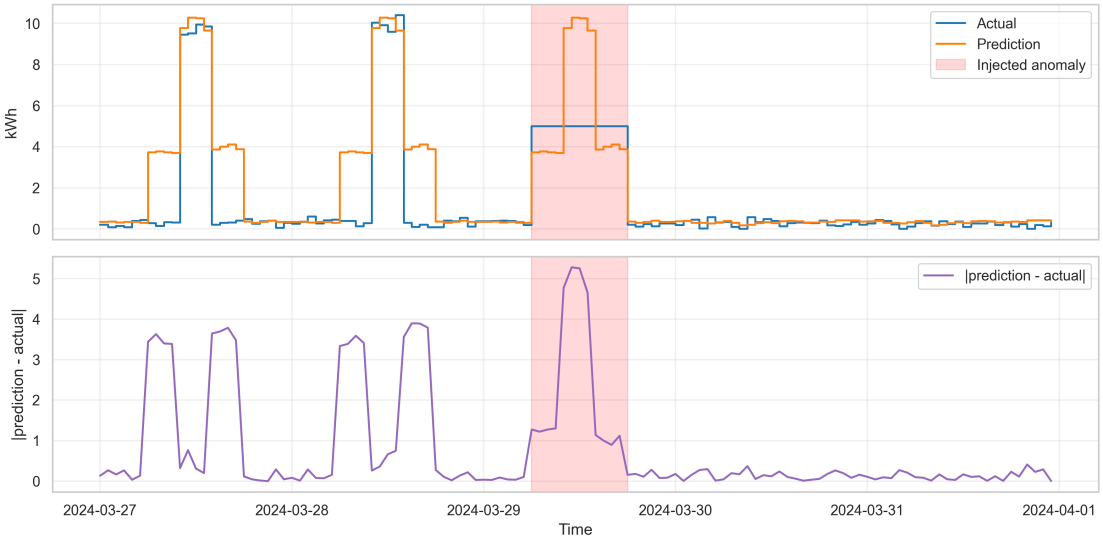


Figure 4.7: Dense regressor residuals over the Variable Shift dataset. The mid-range prediction inflates anomaly scores for legitimate operating states, while the stuck-at-5 fault yields a small residual.

4.6.2. The Gaussian Distribution Paradox

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

4.6.3. Solution: Mixture Density Networks

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

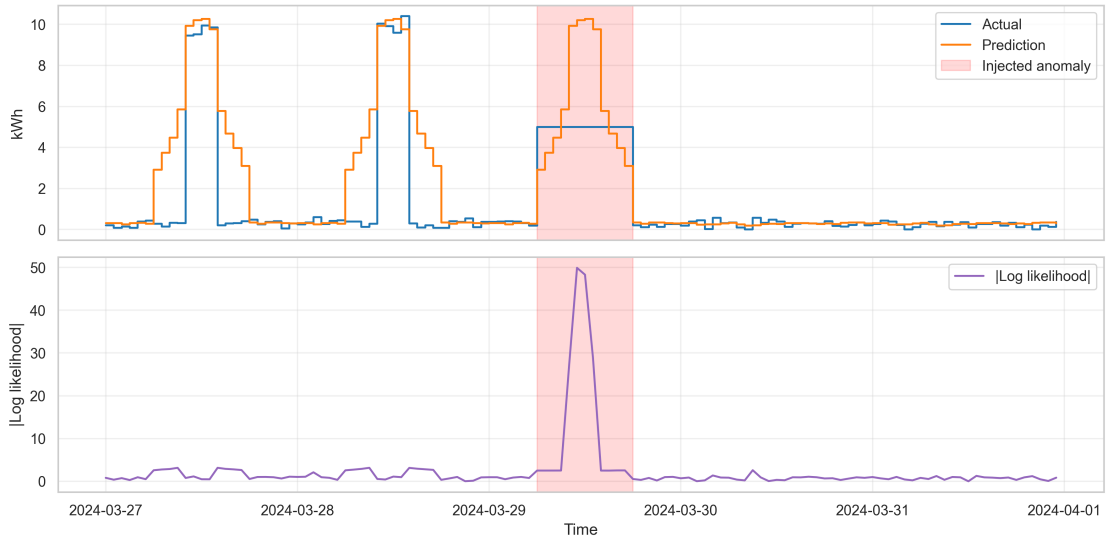


Figure 4.8: Absolute log-likelihood trace for the single-Gaussian predictor. The stuck-at-5 anomaly aligns with the high-likelihood center, suppressing the score.

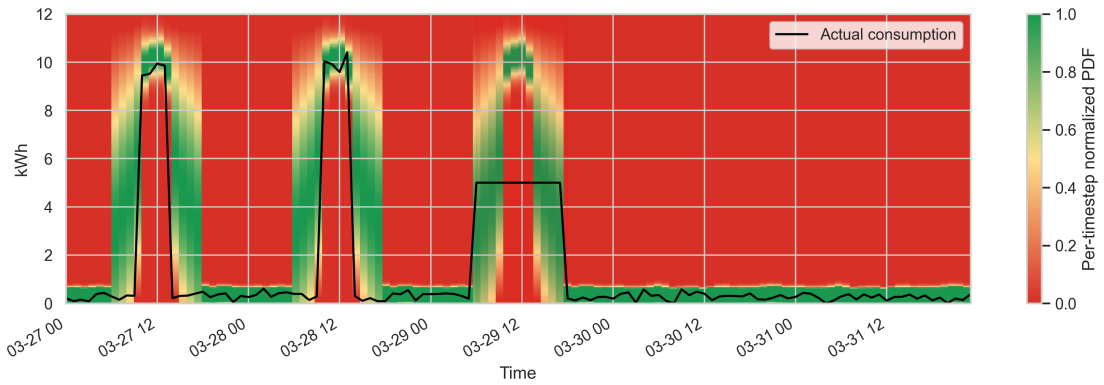


Figure 4.9: Per-timestep normalized PDF for the Gaussian model. High probability mass accumulates between the actual clusters, illustrating the variance-stretching paradox.

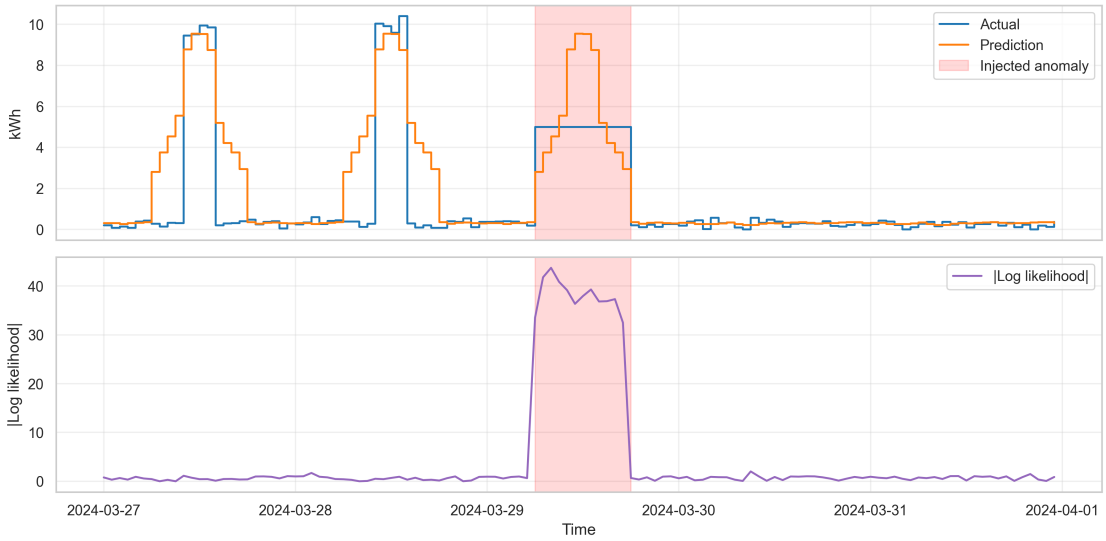


Figure 4.10: MDN absolute log-likelihood trace. The stuck-at-5 anomaly triggers a sustained spike because the value resides in a low-probability region between mixture components.

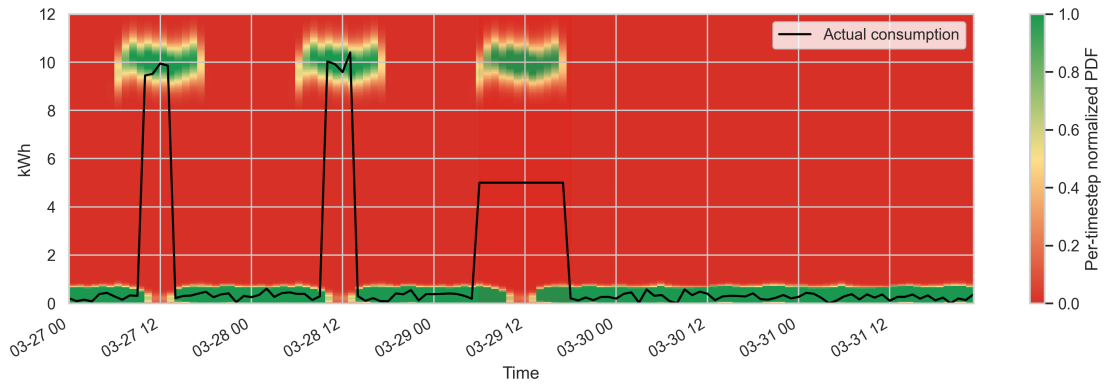


Figure 4.11: MDN normalized PDF heatmap. Two distinct high-probability ridges align with the real operating modes, while the middle band remains improbable.

References

[Su+19] Ya Su et al. “Robust anomaly detection for multivariate time series through stochastic recurrent neural network”. In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 2828–2837.

[Pap+22] John Paparrizos et al. “TSB-UAD: an end-to-end benchmark suite for univariate time-series anomaly detection”. In: *Proceedings of the VLDB Endowment* 15.8 (2022), pp. 1697–1711.

[LP24] Qinghua Liu and John Paparrizos. “The elephant in the room: Towards a reliable time-series anomaly detection benchmark”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 108231–108261.

[Ans+25] Abdul Fatir Ansari et al. “Chronos-2: From univariate to universal forecasting”. In: *arXiv preprint arXiv:2510.15821* (2025).

[Azz+25] Davide Azzalini et al. “An empirical evaluation of deep autoencoders for anomaly detection in the electricity consumption of buildings”. In: *Energy and Buildings* 327 (2025), p. 115069. ISSN: 0378-7788. DOI: https://doi.org/10.1016/j.enbuild.2024.115069. URL: https://www.sciencedirect.com/science/article/pii/S037877882401185X.

[Eli25a] Eliona IoT Platform. *Asset Modeling – Create Templates*. Online documentation. 2025. URL: https://doc.eliona.io/collection/eliona-english/documentation/assets/asset-modeling-create-templates (visited on 12/23/2025).

[Eli25b] Eliona IoT Platform. *Assets*. Online documentation. 2025. URL: https://doc.eliona.io/collection/eliona-english/documentation/assets (visited on 12/23/2025).

[Eli25c] Eliona IoT Platform. *Introduction to Ontologies*. Online documentation. 2025. URL: https://doc.eliona.io/collection/eliona-english/academy/introduction-to-ontologies (visited on 12/20/2025).

[Eli25d] Eliona IoT Platform. *Rule Chains*. Online documentation. 2025. URL: https://doc.eliona.io/collection/eliona-english/documentation/alarms-rules-and-escalations/rule-chains (visited on 12/23/2025).

[Eli25e] Eliona IoT Platform. *Rules*. Online documentation. 2025. URL: https://doc.eliona.io/collection/eliona-english/documentation/alarms-rules-and-escalations/rules (visited on 12/23/2025).

[Eli25f] Eliona IoT Platform. *Structuring Assets*. Online documentation. 2025. URL: https://doc.eliona.io/collection/eliona-english/documentation/assets/structuring-assets (visited on 12/23/2025).

[HHA25] Basu Hela, Praveen Prasad Handigol, and Pandarasamy Arjunan. "Are Time Series Foundation models good for Energy Anomaly Detection?" In: *Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems*. E-Energy '25. Association for Computing Machinery, 2025, pp. 656–665. ISBN: 9798400711251. DOI: 10.1145/3679240.3734633. URL: https://doi.org/10.1145/3679240.3734633.

[MM25] Roya Morshedi and S. Mojtaba Matinkhah. "A Comprehensive Review of Deep Learning Techniques for Anomaly Detection in IoT Networks: Methods, Challenges, and Datasets". In: *Engineering Reports* 7.9 (2025), e70415. DOI: https://doi.org/10.1002/eng2.70415. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/eng2.70415. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/eng2.70415.