ScaloVit-EBM: Localized Energy-Based Anomaly Detection on Time–Frequency Scalograms

Shosuke Asano

32201303

FIT5128: Masters Thesis Final

Supervisor **Dr. Loo Junn Yong**

Co-Supervisor **Dr. Ting Chee Ming**

Word Count for Part 2 (Research Paper): 5556 words

October 2025

Contents

Part 1: Literature Review & Proposal **1**

Part 2: Research Paper **20**

1 Introduction **21**

2 Related work/Background **22**

2.1 Unsupervised Time-Series Anomaly Detection (TSAD) 23

2.2 Signal-to-Image Transformations for Time-Series Data 23

2.3 Energy-Based Models (EBMs) for Anomaly Detection 24

3 Methodology **25**

3.1 Signal Preprocessing and Time–Frequency Representation 25

3.2 Localized Energy-Based Model 26

3.3 Model Training via Energy Matching 27

3.4 Anomaly Detection and Localization 28

4 Experiments and Results **28**

4.1 Dataset Description 28

4.2 Quantitative Comparison with Baselines 29

4.3 Ablation Study 31

4.4 Qualitative Analysis of Anomaly Patterns 33

5 Discussion **34**

5.1 Implications and Insights 34

5.2 Limitations 35

5.3 Future Work 35

6 Conclusion **35**

Acknowledgments **36**

References **36**

Appendix **39**

Part 1: Literature Review

Hybrid Energy-Based Models and Flow Matching for Sensor Anomaly Detection

Literature Review & Project Proposal

(FIT5126)

Shosuke Asano
Student
sasa0005@student.monash.edu

Dr Loo Junn Yong
Supervisor
loo.junnyong@monash.edu

Dr Ting Chee Ming
Co-Supervisor
ting.cheeming@monash.edu

5880 words

Submission Date: 4th May 2025

Table of Contents

1. Introduction 3
2. Literature Review **4**
2.1. Background 4
2.1.1. Common Sensor Fault Types in Anomaly Detection 4
2.1.2. Signal Transformation Techniques 5
2.1.3. Modern Approaches to Sensor Anomaly Detection 6
2.1.4. Generative Models for Anomaly Detection 6
2.2. Related Works 7
2.2.1. Energy-Based Models (EBMs) for Anomaly Detection 7
2.2.2. Diffusion Models for Anomaly Detection 8
2.2.3. Flow Matching for Anomaly Detection 9
2.2.4. Hybrid Approaches (EBMs and Flow Matching) 9
3. Summary of the State of the Art **10**
4. Research Project Plan **11**
4.1. Methodology 11
4.1.1. Data Preparation 12
4.1.2. Model Implementation 13
4.2. Evaluation Strategy 14
4.3. Ethical Considerations 15
5. Conclusion **15**
6. References **16**

1. Introduction

Sensor signals are indispensable components in countless modern systems, providing the data essential for monitoring, control, and automation across industries ranging from manufacturing to healthcare (Zamanzadeh Darban et al., 2024). However, the reliability of these systems hinges on the integrity of sensor data, which can be compromised by various faults or anomalies such as bias, drift, spikes, or stuck values (Li et al., 2020; Yi et al., 2017). These anomalies, arising from sensor degradation, environmental factors, or system malfunctions, can lead to significant operational inefficiencies, safety hazards, and economic losses. Detecting these anomalies is complicated by the nature of contemporary sensor data, often characterized by high dimensionality, noise, complex temporal dependencies, and non-stationarity (Shiva et al., 2024). Consequently, robust anomaly detection methods are critical. While traditional statistical process control techniques have been used (Montgomery, 2024), they often fall short when faced with the complexities and dynamics of modern sensor signals, sometimes requiring unrealistic assumptions about data distributions. The rise of machine learning and deep learning has provided powerful alternatives, particularly unsupervised generative models, which excel at learning complex patterns from unlabeled data (Hoh et al., 2022). These models learn the distribution of normal operational data and identify anomalies as deviations, circumventing the common challenge of scarce labeled anomaly data in real-world settings (Geiger et al., 2020). Key generative approaches include Energy-Based Models (EBMs), Diffusion Models, and Flow Matching.

Given this context, the primary aim of the literature review within this report is to comprehensively survey and synthesize foundational concepts alongside the state-of-the-art in generative modeling techniques as applied to sensor anomaly detection. This involves examining the underlying principles, strengths, and limitations of dominant and emerging methods, assessing their suitability for handling typical sensor data characteristics and fault types, and identifying critical knowledge gaps in current research. To fulfill this aim, the Literature Review (Section 2) first establishes essential background on common sensor faults and signal transformation techniques like STFT (Oppenheim & Lim, 1981) and DST (Dewhurst et al., 2020) crucial for feature extraction. It then critically examines prominent generative models:

- **Energy-Based Models (EBMs):** Valued for their potential interpretability via energy scores but often hampered by intractable partition functions and challenging MCMC-based training (Du & Mordatch, 2019; Nijkamp et al., 2019; Yoon et al., 2023).
- **Diffusion Models:** Known for strong generative performance but typically rely on computationally intensive, iterative reconstruction processes for anomaly detection, often lacking transparency (Liu et al., 2025; Pintilie et al., 2023).
- **Flow Matching:** A more recent framework offering efficient and stable training for continuous normalizing flows, potentially faster than diffusion methods (Lipman et al., 2023; Patel et al., 2024). The review also explores nascent Hybrid Approaches that propose combining EBM principles with Flow Matching training methodologies to potentially achieve both interpretability and efficiency (Balcerak et al., 2025; Loo et al., 2025).

Synthesizing these findings (Section 3), the literature review reveals a significant research gap: despite the theoretical appeal of hybrid EBM/Flow Matching models, there is a lack of empirical studies validating their performance and practical applicability specifically for detecting diverse anomaly types within sensor signal data. Addressing this gap is the core motivation for the research project proposed herein. This project focuses on the development, implementation, and rigorous evaluation of a hybrid EBM/Flow Matching model, employing techniques like Variational Potential

Flow Bayes (VPFB), explicitly tailored for unsupervised sensor anomaly detection. The importance of this work lies in its potential to yield a solution that synergizes the interpretability advantages of EBMs with the computational efficiency and stability of Flow Matching. Successfully demonstrating such a model could lead to more robust, scalable, and trustworthy anomaly detection systems, particularly valuable in domains where understanding the reasoning behind an anomaly flag is crucial (Rudin, 2019). Providing empirical evidence for this approach on benchmark sensor data (MFP dataset, Ruiz-Cárcel et al., 2015) with realistic fault scenarios moves beyond theoretical promise to practical application.

The remainder of this report details this endeavour. Section 2 presents the full literature review, Section 3 summarizes the state-of-the-art and the identified gap, Section 4 outlines the proposed Research Project Plan, and Section 5 offers concluding remarks.

2. Literature Review

This section provides a comprehensive review of approaches to sensor anomaly detection. Through this review, we aim to bridge foundational understanding with cutting-edge approaches, highlighting the evolution from basic principles to novel methodologies.

2.1. Background

This section provides foundational context for sensor anomaly detection by outlining common types of sensor faults, signal transformation techniques, and general approaches to anomaly detection. Together, these components establish a basis for understanding the challenges and motivations behind recent advances in the field.

2.1.1. Common Sensor Fault Types in Anomaly Detection

Anomalies in sensor data are categorized into five common types of sensor faults—bias, drift, erratic, spike, and stuck—as identified in prior studies (Li et al., 2020; Yi et al., 2017). As illustrated in Figure 1, each fault type exhibits distinct characteristics. Bias faults introduce a constant offset to the sensor readings, while drift faults involve a gradual deviation over time. Erratic faults result in unpredictable, noisy fluctuations. Spike faults manifest as sudden, short-lived deviations, and stuck faults occur when the sensor output remains fixed at a constant value. These faults persist throughout the entire input sequence, with each component of the input equally likely to be affected.

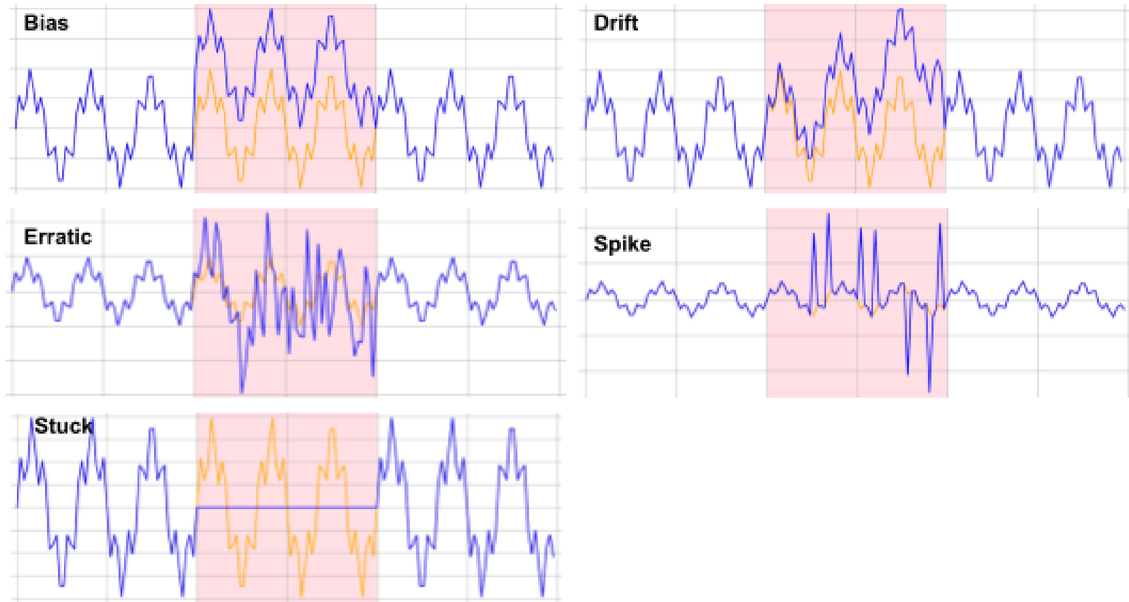


Figure 1: Examples of sensor faults (blue line) compared to the correct signal (orange line).

2.1.2. Signal Transformation Techniques

Sensor signals often exhibit complex, non-stationary temporal dynamics that are not easily captured in their raw time-domain form. To better extract meaningful features, signal transformation techniques are commonly applied prior to modeling (Zhao et al., 2018).

One widely used method is the Short-Time Fourier Transform (STFT), which converts the signal into a time-frequency representation (Oppenheim & Lim, 1981) as illustrated in Figure 2. STFT divides the signal into overlapping segments and applies the Fourier transform to each window, enabling the extraction of localized frequency components over time. This representation allows models to detect transient events, periodic structures, and evolving spectral patterns, all of which may indicate anomalies. Recent studies have shown the effectiveness of STFT-based features in generative models for time series prediction, especially in contexts with strong temporal dependencies and non-stationary behavior (Naiman et al., 2024). By expanding the feature space, STFT enhances a model's ability to learn complex patterns and improves its capacity to detect subtle anomalies.

Complementing frequency-based approaches, the Discrete Shocklet Transform (DST) provides a shape-based method for signal transformation (Dewhurst et al., 2020). Unlike STFT, DST does not rely on frequency decomposition but instead uses cross-correlation with predefined "shock-like" kernels to identify localized patterns in the time domain. This approach is timescale-independent and focuses on detecting abrupt changes and transient behaviors that do not exhibit clear periodicity. DST is particularly valuable in applications where anomalies manifest as sudden, irregular shapes rather than frequency shifts, offering a mechanism-driven, interpretable view of the data.

Together, STFT and DST provide complementary perspectives on signal dynamics—STFT capturing oscillatory and spectral structures, and DST highlighting shape-based, non-oscillatory patterns—both enriching the input representation for downstream anomaly detection models.

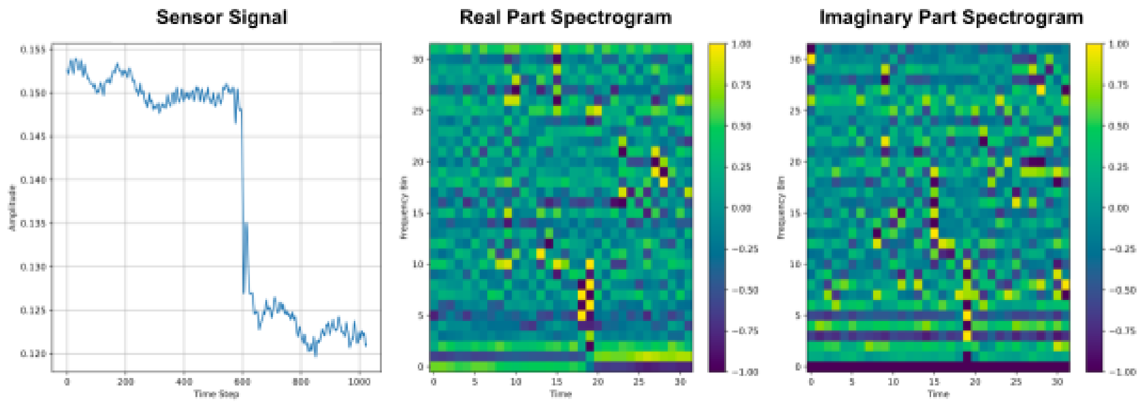


Figure 2: Real and imaginary components of the spectrograms generated by applying the Short-Time Fourier Transform (STFT) to the corresponding signal.

2.1.3. General Approaches to Anomaly Detection

Anomaly detection in sensor data is a well-established field with a rich history spanning several decades. It plays a critical role across various domains, including fraud detection, cybersecurity, network monitoring, industrial systems, and healthcare (Zamanzadeh Darban et al., 2024). Traditional approaches to outlier detection have often relied on statistical methods, which typically assume specific underlying data distributions and methods that depend on predefined thresholds (Montgomery, 2024). However, contemporary sensor data presents significant challenges to these traditional techniques due to its complex characteristics, including high dimensionality, inherent noise, temporal dependencies, and the phenomenon of concept drift, where the statistical properties of the data evolve over time (Shiva et al., 2024).

To overcome the limitations of traditional methods, recent research has increasingly focused on machine learning and, more notably, deep learning-based approaches for anomaly detection (Zamanzadeh Darban et al., 2024). These methods excel at capturing complex, nonlinear patterns in data and do not require strong assumptions about the underlying distribution. Deep learning models, in particular, are well-suited for handling high-dimensional data such as multivariate time series (Zamanzadeh Darban et al., 2024), making them highly effective in modern sensor-based applications. However, a common drawback of deep neural networks is their lack of interpretability, which can be a major limitation in domains such as finance and healthcare, where transparency and explainability are critical (Montgomery, 2024; Rudin, 2019).

2.1.4. Generative Models for Anomaly Detection

Among modern techniques, generative models have emerged as a particularly promising direction for anomaly detection in time series data (Hoh et al., 2022). These models learn the underlying distribution of normal patterns and detect anomalies by identifying data points with low likelihood or high reconstruction error under the learned distribution. This paradigm enables more robust detection of subtle and complex anomalies compared to traditional rule-based systems. In addition to their effectiveness, generative models offer several practical advantages, especially in real-world, unsupervised settings (Geiger et al., 2020):

- **No need for labeled anomalies:** Generative models operate in an unsupervised manner, learning directly from raw time series data without requiring prior knowledge of known anomalies.

- **Independence from simulation-based "normal baselines":** They do not rely on physics-based models or predefined normal signals, making them suitable for systems where generating or defining a normal baseline is difficult.
- **Adaptability to real-world noise and external variability:** Some generative models, such as flow matching (Lipman et al., 2023), exhibit robustness to data containing unexpected shifts or patterns caused by external phenomena (e.g., environmental changes or control regime shifts).
- **Applicability to unsegmented and variable-length data:** Generative approaches are well-suited for signals that cannot be easily segmented or have variable durations, which is common in many industrial and sensor-based applications.

By combining strong modeling capacity with adaptability and minimal supervision requirements, generative models stand out as a scalable solution for anomaly detection in complex, real-world environments.

2.2. Related Works

Recent generative modeling advances have opened new avenues for unsupervised sensor anomaly detection, particularly in settings with scarce labeled data. While originally developed for tasks like image synthesis or density estimation, models such as Energy-Based Models (EBMs), Diffusion Models, and Flow Matching have shown strong potential for capturing complex data distributions and identifying anomalies. This section reviews these approaches with a focus on their applicability to anomaly detection and highlights emerging hybrid models that aim to combine the interpretability of EBMs with the efficiency of Flow Matching.

2.2.1. Energy-Based Models (EBMs) for Anomaly Detection

Energy-Based Models (EBMs) offer a powerful and flexible framework for modeling complex, high-dimensional data distributions. At their core, EBMs define an energy function E(x), typically parameterized by a neural network, which assigns low energy values to in-distribution (normal) data points and higher values to out-of-distribution (anomalous) ones (Nijkamp et al., 2019; Yoon et al., 2023). Since anomalies often lie in low-density regions of the data manifold, they are effectively characterized by high energy scores (Du & Mordatch, 2019).

In the context of anomaly detection, the key inference task involves computing the energy E(x) for a given test input. A sample is flagged as anomalous if its energy exceeds a predefined threshold (Yoon et al., 2023). While finding configurations that minimize the energy function is a related inference task within EBMs (used in training or generation), direct energy evaluation is the standard approach for anomaly scoring. Probabilistically, EBMs define a Boltzmann distribution over data x as:

$$P(x) = \frac{e^{-E(x)}}{Z}$$

Here, E(x) represents the energy assigned to data point x, and $Z = \int e^{-E(x)} dx$ is the partition function. It is an intractable normalization constant for high-dimensional data. Obtaining samples directly from this distribution is also generally intractable. This intractability poses the primary challenge for maximum likelihood training of EBMs (Xiao et al., 2021). To address this, sampling methods, particularly Markov Chain Monte Carlo (MCMC) techniques, are frequently employed to approximate the required expectations. Consequently, several approximate training methods have been developed to circumvent this challenge:

- Contrastive Divergence (CD): Approximates the log-likelihood gradient using a small number of sampling steps (e.g., Langevin dynamics or Gibbs sampling) starting from observed data (Du et al., 2020).
- Score Matching (SM): Avoids the partition function altogether by minimizing the difference between the score functions (gradients of log-probability) of the model and the empirical data distribution (Li et al., 2023).
- Noise Contrastive Estimation (NCE): Reframes density estimation as a binary classification task, distinguishing real data from noise samples drawn from a known distribution (Ma & Collins, 2018).

As discussed in Section 2.1.3, model interpretability is a critical consideration, especially in high-stakes domains like finance and healthcare. In this regard, EBMs offer a degree of transparency by directly modeling energy landscapes that are semantically meaningful (Carbone, 2024). Furthermore, their unsupervised training framework aligns naturally with real-world anomaly detection scenarios, where labeled anomalies are rare or unavailable.

While the training of EBMs remains a significant challenge due to the intractability of the partition function and the complexity of sampling-based methods, which often involve computationally expensive MCMC procedures with potential convergence issues, ongoing advances in approximate learning techniques continue to enhance their viability for practical use. Given their expressive power, interpretability, and compatibility with unsupervised learning, EBMs represent one of the compelling approaches for robust, scalable anomaly detection.

2.2.2. Diffusion Models for Anomaly Detection

Diffusion Models are a powerful class of generative models that have achieved remarkable performance across various domains, including image synthesis, speech generation, and time-series modeling (LingYang et al., 2023). Their core mechanism involves a two-step process: a forward diffusion process, where noise is gradually added to the training data over several time steps, and a reverse diffusion process, where the model learns to denoise this noisy data to recover the original input distribution. The reverse process is typically parameterized by a neural network (often a U-Net or a Transformer), which is trained to predict and remove the noise introduced during the forward process at each time step.

In the field of anomaly detection, diffusion models are primarily used in reconstruction-based frameworks (Pintilie et al., 2023). These approaches involve training the diffusion model exclusively on normal data. During inference, the model attempts to reconstruct potentially anomalous test inputs by running the learned reverse diffusion process. If the input is anomalous, the model typically struggles to reconstruct it accurately. The reconstruction error, measured in terms of the difference between the input and the generated output, serves as the anomaly score (Pintilie et al., 2023). Larger errors indicate a higher likelihood of anomaly. Beyond reconstruction error, recent studies have also investigated alternative anomaly scoring methods. Some utilize the learned latent representations or directly leverage the score function produced during training as an indicator of deviation from the learned distribution (Sakai & Hasegawa, 2025). These methods aim to exploit deeper internal signals from the generative process for more sensitive or interpretable detection.

Despite their strong modeling capacity, diffusion models face notable practical challenges, particularly in terms of computational cost. Because the reverse denoising process involves hundreds to thousands of iterative steps, computing an anomaly score at inference time can be time-consuming, especially when applied to long sequences or high-dimensional data (Liu et al., 2025; Yang et al., 2023). This makes diffusion-based anomaly detection less suitable for real-time or resource-constrained scenarios.

Compared to Energy-Based Models (EBMs) discussed in Section 2.2.1, diffusion models tend to offer greater training stability and end-to-end differentiability, avoiding the need for MCMC-based inference techniques that are often sensitive to hyperparameter tuning and can be slow to converge (Carbone, 2024). However, like many deep generative models, diffusion models also lack interpretability (Katsuoka et al., 2024; Liu et al., 2025), making it difficult to provide clear explanations for why certain inputs are classified as anomalous. This lack of transparency can hinder their adoption in high-stakes, real-world settings where trust and explainability are critical.

2.2.3. Flow Matching for Anomaly Detection

Flow Matching is a recent and efficient generative modeling framework designed to train Continuous Normalizing Flows (CNFs) without relying on simulation-based solvers (Lipman et al., 2023). CNFs transform a simple base distribution into complex data distributions through a continuous-time flow defined by a time-dependent vector field. Traditional CNF training requires solving differential equations, which can be slow and unstable. Flow Matching simplifies this by directly regressing a neural network onto a target vector field, which is defined by interpolated probability paths between noise and data samples.

A major strength of Flow Matching lies in its training efficiency and stability. Unlike diffusion models, which rely on second-order partial differential equations (PDEs) and require hundreds of denoising steps, Flow Matching trains using first-order PDEs, resulting in faster convergence and more robust learning (Patel et al., 2024). In terms of inference speed, Flow Matching often produces straighter and more direct sampling paths, which reduce the number of steps needed to evaluate or generate data samples compared to diffusion-based models (Lipman et al., 2023).

Furthermore, its flexibility allows it to integrate with techniques from diffusion models and path-based generative methods, offering a unified and generalizable framework for generative modeling (Lipman et al., 2023). It can also be used to train EBMs, offering better training stability and efficiency compared to MCMC-based methods (Chao et al., 2023). This makes Flow Matching a promising and adaptable tool in the growing landscape of unsupervised anomaly detection, especially when balancing modeling power, interpretability, and computational efficiency.

2.2.4. Hybrid Approaches (EBMs and Flow Matching)

As previously discussed in Section 2.2.1, EBMs offer a highly interpretable framework for unsupervised anomaly detection by assigning scalar energy values to data points—low energy to in-distribution data and high energy to anomalies. Despite their theoretical appeal and compatibility with label-scarce scenarios, EBMs suffer from practical limitations, most notably the intractability of the partition function and the reliance on slow, often unstable MCMC methods for sampling and training (Du & Mordatch, 2019).

In contrast, Flow Matching (Section 2.2.3) has emerged as a training-efficient, simulation-free alternative for modeling data distributions through CNFs. By learning a deterministic vector field between base and target distributions via regression, rather than solving differential equations or relying on iterative denoising, it supports stable, first-order PDE-based training and faster inference (Patel et al., 2024). Moreover, Flow Matching provides a versatile generative framework that aligns well with both diffusion models and path-based generative techniques, offering a bridge between model expressiveness and computational feasibility (Lipman et al., 2023).

Recent hybrid approaches propose leveraging the strengths of both EBMs and Flow Matching to address the shortcomings of each (Loo et al., 2025). Specifically, they aim to retain the interpretability and outlier sensitivity of EBMs while improving training and inference scalability through Flow Matching. A central idea in these methods is to train a Flow Matching model such that its induced distribution converges to a Boltzmann distribution. This allows the EBM's energy function to be implicitly learned within the parameters of the flow model, bypassing the need for explicit energy function parameterization and traditional MCMC-based training (Loo et al., 2025).

One such approach, rooted in the Jordan–Kinderlehrer–Otto (JKO) scheme for Wasserstein gradient flows, proposes learning a scalar potential V whose corresponding Boltzmann distribution models the data (Balcerak et al., 2025). This is typically achieved via a two-phase training strategy: an initial warm-up phase using a flow-like objective at zero temperature ($\varepsilon=0$) to efficiently transport noise samples towards the data manifold and generate high-quality negative samples, followed by a main phase that jointly optimizes the flow objective with a contrastive divergence loss at increasing temperatures (ε) to refine the potential and ensure the learned Boltzmann distribution accurately reflects the data's likelihood (Balcerak et al., 2025).

For anomaly detection, the implicitly learned energy landscape can be directly used to assign energy scores to new data points. As with traditional EBMs, those with high energy (low likelihood under the model) are flagged as anomalies (Yoon et al., 2023). Crucially, this approach eliminates the slow MCMC-based sampling loops typically required for inference in EBMs, making it scalable and robust even in high-dimensional settings (Loo et al., 2025).

3. Summary of the State of the Art

Sensor anomaly detection is a critical task challenged by the complex, high-dimensional, and dynamic nature of modern sensor data (Shiva et al., 2024; Zamanzadeh Darban et al., 2024). While traditional statistical methods struggle with these complexities (Montgomery, 2024), machine learning, particularly deep generative models, has emerged as a powerful alternative capable of learning intricate patterns in an unsupervised manner (Hoh et al., 2022). These models learn the distribution of normal data and identify anomalies as deviations, overcoming the need for labeled anomaly data, which is often scarce (Geiger et al., 2020). Among generative approaches, several prominent models have been explored. EBMs provide an interpretable framework by assigning an energy score to data points, with higher scores indicating anomalies (Nijkamp et al., 2019; Yoon et al., 2023). Their main drawback lies in challenging training dynamics, often requiring computationally expensive and potentially unstable MCMC methods due to the intractable partition function (Du & Mordatch, 2019; Xiao et al., 2021). Diffusion Models represent another potent class, achieving state-of-the-art generation quality (LingYang et al., 2023). However, their application to anomaly detection typically relies on reconstruction error, which can be computationally intensive due to the iterative denoising process, and they suffer from a lack of interpretability (Katsuoka et al., 2024; Liu et al., 2025).

More recently, Flow Matching has gained attention as an efficient and stable method for training CNFs without simulation (Lipman et al., 2023). It offers faster training and potentially faster inference compared to diffusion models, using direct regression onto target vector fields (Lipman et al., 2023; Patel et al., 2024). Flow Matching also presents flexibility, bridging concepts from diffusion and path-based generation, and has shown potential for stabilising EBM training (Chao et al., 2023; Lipman et al., 2023). This has led to emerging hybrid approaches that aim to combine the interpretability of EBMs with the training efficiency and stability of Flow Matching (Loo et al., 2025). These methods propose learning the EBM's energy function implicitly within a Flow Matching framework, potentially leveraging techniques like the JKO scheme to bypass traditional EBM training hurdles (Balcerak et al., 2025; Loo et al., 2025). The literature describes how the learned energy landscape can be used for anomaly detection by scoring data points (Yoon et al., 2023).

However, a clear gap exists in the empirical validation and specific application of these novel hybrid EBM and Flow Matching models directly for the task of sensor signal anomaly detection. While the theoretical foundations, improved performance and potential benefits of combining these frameworks have been proposed (Balcerak et al., 2025; Loo et al., 2025), their practical effectiveness, scalability, and performance compared to established methods within the specific context of the anomaly detection for complex signal data (characterized by phenomena like bias, drift, and spikes) remain unexplored in the current researches. The proposed research project aims to directly fill this gap. It will involve developing, implementing, and rigorously evaluating a hybrid model leveraging Flow Matching to train an EBM specifically tailored for sensor signal anomaly detection. This project will investigate whether this hybrid approach can effectively retain the interpretability benefits of EBMs while harnessing the computational efficiency of Flow Matching, thereby providing a robust, scalable, and interpretable solution for identifying diverse fault types in real-world sensor signals. The evaluation will focus on performance metrics and the model's ability to handle various sensor fault types compared to existing generative approaches.

4. Research Project Plan

This section outlines the plan to address the identified gap in Section 3: the lack of empirical validation and specific application of hybrid EBMs and Flow Matching models for the task of sensor signal anomaly detection. The plan details the objectives, methodology, evaluation strategy, and ethical considerations for developing and assessing such a hybrid model.

The core aim is to empirically investigate the practical possibility of combining EBMs and Flow Matching for robust, interpretable, and efficient sensor anomaly detection. This translates into the following specific objectives:

1. **Develop a Hybrid EBM/Flow Matching Model for Sensor Signal Anomaly Detection:** Implement a functional hybrid model inspired by recent theoretical work (e.g., Balcerak et al., 2025; Loo et al., 2025).
2. **Train and Test the Model on Benchmark Sensor Data:** Employ the dataset, training the model in an unsupervised fashion. Generate realistic fault types for robust anomaly detection evaluation.
3. **Evaluate Performance:** Assess the model’s performance using standard anomaly detection metrics to determine its effectiveness.

4.1. Methodology

In this section, we describe the full pipeline for anomaly detection, starting with signal preprocessing—normalization, segmentation, and transformation—to extract meaningful temporal and spectral features. We then implement and train a hybrid Energy-Based Model with Flow Matching on clean sensor data to learn its normal distribution, enabling unsupervised detection of anomalies via learned energy scores.

4.1.1. Data Preparation

We utilize the Multiphase Flow Process (MFP) dataset introduced by Ruiz-Cárcel et al. (2015), which captures time-series data from a wide range of industrial-grade sensors monitoring a controlled flow of water, oil, and air through a complex pipeline system. The dataset captures rich, nonlinear sensor dynamics characteristic of real-world industrial processes, making it a valuable resource for developing and evaluating anomaly detection models under realistic operating conditions.

The data preprocessing pipeline consists of three key phases: normalization, segmentation, and signal transformation. Each phase plays an important role in preparing the raw sensor signals for effective learning and anomaly detection.

- **Normalization:** Sensor signals can vary significantly in scale depending on the type of sensor or measurement unit used. For example, one sensor may produce values ranging from 0 to 100, while another may range from 0 to 1. To ensure consistent input across different sensors and to stabilize training, each signal is normalized to a common scale. This helps the model focus on the signal's shape and pattern rather than absolute magnitudes.
- **Segmentation:** Sensor data is typically continuous and high-frequency. To process this data with models that expect fixed-size input, the signals are segmented into overlapping windows of a fixed length. This converts the time series into a sequence of smaller sub-sequences, where each window captures a local pattern of signal behavior. The window size and overlap ratio are treated as hyperparameters that can be tuned based on preliminary experiments. This windowing approach also provides a temporal context for anomaly detection.
- **Signal Transformation:** To enhance the representation of each segmented window, signal transformation techniques are applied. These transformations extract meaningful features that reveal both the temporal and spectral properties of the signal.
    - The Short-Time Fourier Transform (STFT) (Oppenheim & Lim, 1981) is used to convert each time-domain window into a time-frequency representation, as illustrated in Figure 2. STFT helps the model capture non-stationary behaviors such as drifts, spikes, or frequency shifts, which are common in sensor faults.
    - In addition, the Discrete Shocklet Transform (DST) (Dewhurst et al., 2020) may be explored to extract shape-based features from the signals. DST is particularly effective at capturing sharp transitions and localized signal changes, complementing the smooth, frequency-focused representations from STFT. The inclusion of DST will be based on exploratory analysis and its contribution to model performance.

For evaluation purposes, we generate labeled test samples by injecting five common sensor fault types—bias, drift, erratic, spike, and stuck—into segments of the normal MFP dataset, following augmentation methodologies proposed in prior work (Ding et al., 2022). To simulate both in-distribution and out-of-distribution fault scenarios, we introduce positive bias and drift, as well as normally distributed erratic noise. Additionally, we inject negative bias and drift, uniformly distributed erratic noise, and two distinct fault types—spike and stuck—that represent more challenging, unseen conditions. These synthetic faults are applied exclusively during testing to rigorously evaluate model robustness, while the model itself is trained in an unsupervised manner on clean, fault-free data. This setup enables a comprehensive assessment of anomaly detection performance across both familiar and novel fault distributions.

4.1.2. Model Implementation

The core of our generative model is built upon the principles of Variational Potential Flow Bayes (VPFB), drawing inspiration from recent advancements (Loo et al., 2025). Instead of directly training a vector field to match predefined targets, this approach focuses on learning a time-dependent potential energy function $\Phi(x, t)$, parameterized by a neural network. The negative gradient of this potential, $-\nabla_x \Phi(x, t)$, implicitly defines the velocity field of a continuous-time flow. The model is trained such that the probability density transported by this potential flow matches a target density homotopy (a smooth path of distributions from a simple prior to the data distribution) over time $t \in [0, t_{end}]$. A key theoretical result of VPFB is that, as the density homotopy approaches a stationary equilibrium (for $t \geq t_{max}$), the learned potential $\Phi(x, t)$ relates directly to the Boltzmann energy $E(x)$ of a standard EBMs, such that $P(x) \propto e^{-E(x)}$. Our goal is to learn this implicit EBM representing the distribution of normal sensor data without relying on traditional, often unstable, MCMC-based EBM training. Figure 3 illustrates the complete pipeline for our VPFB-based sensor signal anomaly detection system, divided into training (upper panel) and anomaly detection (lower panel).

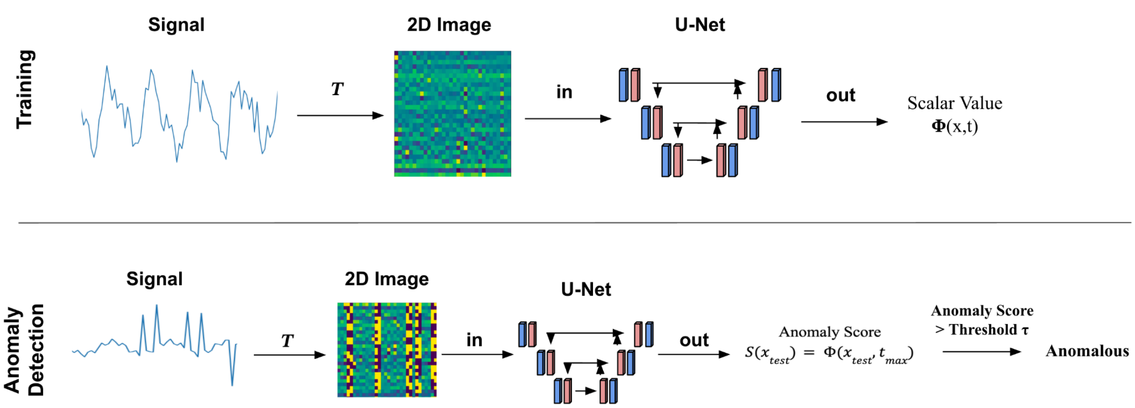


Figure 3: Anomaly Detection Pipeline using Hybrid EBM/Flow Matching. Illustrates training and detection workflows including signal processing, U-Net, and energy scoring.

- **Neural Network Architecture:** Based on VPFB (Loo et al., 2025), the time-dependent potential energy function $\Phi(x, t)$ is parameterized by a neural network based on the U-Net architecture (Ronneberger et al., 2015). Hyperparameters for the U-Net were adopted from Lipman et al. (2023). The network accepts the data input x and time t, outputting the scalar potential energy value $\Phi(x, t)$.

- **Training Protocol:** The neural network parameterizing Φ is trained by minimizing a variational loss function derived from the principle of matching the flow-driven density homotopy to the target marginal density homotopy. This loss, specific to the VPFB framework (Loo et al., 2025), avoids explicit density evaluation and MCMC sampling. The overall training objective is to minimize the expected value of a time-dependent loss $L(\Phi, t)$ over the training duration, effectively averaging over sampled times:

$$min_{\theta} E_{t \sim U(0, t_{end})}[L(\Phi_{\theta}, t)]$$

where Φ_θ denotes the potential function parameterized by network weights θ, and $L(\Phi, t)$ is calculated for each sampled time t and batch of data. The time-dependent loss function is (Loo et al., 2025, Equation 25):

$$L(\Phi, t) = Cov_{p(x|\bar{x},t)P_{data}(\bar{x})}[\Phi(x, t), w(t)\gamma(x, \bar{x}, t)]$$
$$- \frac{\nabla_x \Phi(x,t) \cdot v(x|\bar{x},t)}{||\nabla_x \Phi(x,t)|| \, ||v(x|\bar{x},t)||} + E_{p(x|\bar{x},t)P_{data}(\bar{x})}[||\nabla_{(x,t)} \Phi(x, t)||^2 + \eta||\Phi(x, t)||^2]$$

In practice, the expectations and covariances are computed using mini-batches of samples. For each training step:

1. Sample a batch of normal sensor data segments $\bar{x} \sim P_{data}$.
2. Sample a batch of time points $t \sim U(0, t_{end})$
3. For each $(\bar{x}, t)$ pair, sample a perturbed point x from the conditional distribution $p(x|\bar{x},t)$.
4. Compute the corresponding innovation term $\gamma(x, \bar{x},t)$ and the target conditional vector field $v(x|\bar{x},t)$
5. Compute $\Phi(x, t)$, its spatial gradient $\nabla_x \Phi(x, t)$, and its time derivative $\vartheta_t \Phi(x, t)$ using the neural network and automatic differentiation.
6. Calculate the loss $L(\Phi, t)$ for each sample in the batch and the average.
7. Backpropagate and update network weights using optimizers like Adam.

This training protocol ensures that the learned potential function guides the flow to match the desired density evolution, thereby learning the energy landscape of the normal data distribution. Elements from advanced training strategies, such as multi-stage training based on insights from Balcerak et al. (2025), may be explored to further stabilize training or enhance the discriminative power of the learned energy function. Hyperparameters will be tuned using a validation set of normal data.

- **Anomaly Scoring Mechanism:** Upon successful training of the potential function $\Phi(x, t)$, the implicitly learned energy function for a test sample x_{test} is related to $\Phi(x_{test}, t)$ at or near the stationary equilibrium time t_{max}. Following common practice in energy-based anomaly detection (Yoon et al., 2023), we use the learned potential energy evaluated at this equilibrium time as the anomaly score for a test sample x_{test}:

$$S(x_{test}) = \Phi(x_{test}, t_{max})$$

A higher energy value $S(x_{test})$ indicates that the sample x_{test} falls in a region of low probability density under the learned model of normal data, thus signifying a higher likelihood of being anomalous. Anomalies are identified by setting a threshold on this score; test samples with $S(x_{test})$ exceeding the threshold are flagged as anomalies.

4.2. Evaluation Strategy

To evaluate the effectiveness of our anomaly detection model on time-series sensor data, we adopt the Area Under the Receiver Operating Characteristic Curve (AUROC) as the primary performance metric, following the approach introduced by Koizumi et al. (2020). AUROC is widely used in anomaly detection tasks as it provides a threshold-independent measure of a model's ability to separate normal from anomalous signal patterns, which is particularly important when ground-truth labels are sparse or uncertain. In addition, we report F1-score, precision, and recall, which are standard metrics for evaluating classification performance in imbalanced settings. These metrics allow us to quantify the trade-off between false positives and false negatives, especially relevant in sensor-based monitoring where anomalies are rare but critical to detect. This combination of metrics provides a robust and interpretable evaluation framework for assessing both detection accuracy and reliability.

4.3. Ethical Considerations

The dataset used in this study (Ruiz-Cárcel et al., 2015) contains only industrial sensor data with no human subjects or sensitive information. Its use is permitted under a permissive open-source license, and we have obtained the necessary rights for research use. Thus, there is no need for ethical considerations in this research.

5. Conclusion

This report has provided a comprehensive review of sensor anomaly detection, highlighting the complexities of modern sensor data and the evolution from traditional statistical methods to advanced deep learning approaches. We examined key generative models, including Energy-Based Models (EBMs), valued for their interpretability but challenged by training complexities involving MCMC methods; Diffusion Models, recognized for state-of-the-art performance but often hampered by slow inference and limited explainability; and Flow Matching, an emerging efficient and stable framework for training continuous normalizing flows. Recent theoretical developments propose hybrid models combining the interpretability strengths of EBMs with the computational efficiency and stability of Flow Matching, by learning an implicit energy potential via flow-based objectives like the Variational Potential Flow Bayes (VPFB) framework. However, a notable gap exists in the literature regarding the specific application and empirical validation of these promising hybrid approaches for the distinct challenges of sensor anomaly detection. This research project aims to fill this crucial gap. We propose the development, implementation, and rigorous evaluation of a hybrid EBM/Flow Matching model tailored for sensor signals, utilizing the VPFB approach on the benchmark MFP dataset with synthetically generated faults. The primary objective is to empirically assess if this hybrid strategy can deliver a robust, scalable, and interpretable solution for identifying diverse sensor anomalies, thereby advancing the state-of-the-art in unsupervised anomaly detection for critical monitoring systems.

6. References

Balcerak, M., Amiranashvili, T., Shit, S., Terpin, A., Kaltenbach, S., Koumoutsakos, P., & Menze, B. (2025). *Energy Matching: Unifying Flow Matching and Energy-Based Models for Generative Modeling*. https://doi.org/10.48550/arXiv.2504.10612

Carbone, D. (2024). *Hitchhiker's guide on Energy-Based Models: a comprehensive review on the relation with other generative models, sampling and statistical physics*. https://doi.org/10.48550/arXiv.2406.13661

Chao, C.-H., Sun, W.-F., Hsu, Y.-C., Kira, Z., & Lee, C.-Y. (2023). *Training Energy-Based Normalizing Flow with Score-Matching Objectives*. https://doi.org/10.48550/arxiv.2305.15267

Dewhurst, D. R., Alshaabi, T., Kiley, D., Arnold, M. V., Minot, J. R., Danforth, C. M., & Dodds, P. S. (2020). The shocklet transform: a decomposition method for the identification of local, mechanism-driven dynamics in sociotechnical time series. *EPJ Data Science*, *9*(1). https://doi.org/10.1140/epjds/s13688-020-0220-x

Ding, Z. Y., Loo, J. Y., Surya Girinatha Nurzaman, Tan, C. P., & Baskaran, V. M. (2022). A Zero-Shot Soft Sensor Modeling Approach Using Adversarial Learning for Robustness Against Sensor Fault. *IEEE Transactions on Industrial Informatics*, *19*(4), 5891–5901. https://doi.org/10.1109/tii.2022.3187708

Du, Y., Li, S., Tenenbaum, J., & Igor Mordatch. (2020). Improved Contrastive Divergence Training of Energy Based Models. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2012.01316

Du, Y., & Mordatch, I. (2019). Implicit Generation and Modeling with Energy Based Models. *Neural Information Processing Systems*, *32*, 3603–3613.

Geiger, A., Liu, D., Alnegheimish, S., Cuesta-Infante, A., & Veeramachaneni, K. (2020, December 1). *TadGAN: Time Series Anomaly Detection Using Generative Adversarial Networks*. IEEE Xplore. https://doi.org/10.1109/BigData50022.2020.9378139

Hoh, M., Schöttl, A., Schaub, H., & Wenninger, F. (2022). A Generative Model for Anomaly Detection in Time Series Data. *Procedia Computer Science*, *200*, 629–637. https://doi.org/10.1016/j.procs.2022.01.261

Katsuoka, T., Shiraishi, T., Miwa, D., Nguyen Le Duy, V., & Takeuchi, I. (2024). *Statistical Test on Diffusion Model-based Anomaly Detection by Selective Inference*. https://doi.org/10.48550/arXiv.2402.11789

Koizumi, Y., Kawaguchi, Y., Imoto, K., Nakamura, T., Nikaido, Y., Tanabe, R., Purohit, H., Suefusa, K., Endo, T., Yasuda, M., & Harada, N. (2020). *Description and Discussion on DCASE2020 Challenge Task2: Unsupervised Anomalous Sound Detection for Machine Condition Monitoring*. https://doi.org/10.48550/arXiv.2006.05822

Li, D., Wang, Y., Wang, J., Wang, C., & Duan, Y. (2020). Recent advances in sensor fault diagnosis: A review. *Sensors and Actuators A: Physical*, *309*, 111990.

https://doi.org/10.1016/j.sna.2020.111990

Li, Z., Chen, Y., & Sommer, F. T. (2023). Learning Energy-Based Models in High-Dimensional Spaces with Multiscale Denoising-Score Matching. *Entropy*, *25*(10), 1367–1367. https://doi.org/10.3390/e25101367

LingYang, L., Zhang, Z., Song, Y., Shenda, H., Xu, R., Zhao, Y., Zhang, W., Cui, B., & Yang, M.-H. (2023). Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Computing Surveys*, *56*(4), 1–39. https://doi.org/10.1145/3626235

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., & Le, M. (2023). *Flow Matching for Generative Modeling*. https://doi.org/10.48550/arXiv.2210.02747

Liu, J., Ma, Z., Wang, Z., Liu, Y., Wang, Z., Sun, P., Song, L., Hu, B., Boukerche, A., & C.M. Leung, V. (2025). A Survey on Diffusion Models for Anomaly Detection. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2501.11430

Loo, J. Y. L. P., Vishnu Monn Baskaran, Chee-Ming Ting, Raphaël C.-W. Phan, Adeline, M., Lau, J. K., Leong, F. Y., Tew, H. H., Pal, A., Baskaran, V. M., Ting, C.-M., & Phan, R. C.-W. . (2025). *Learning Energy-Based Generative Models via Potential Flow: A Variational Principle Approach to Probability Density Homotopy Matching*. https://doi.org/10.48550/arXiv.2504.16262

Ma, Z., & Collins, M. (2018). Noise Contrastive Estimation and Negative Sampling for Conditional Models: Consistency and Statistical Efficiency. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3698–3707). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1809.01812

Montgomery, R. M. (2024). *Techniques for Outlier Detection: A Comprehensive View*. https://doi.org/10.20944/preprints202410.1603.v1

Naiman, I., Berman, N., Pemper, I., Arbiv, I., Fadlon, G., & Azencot, O. (2024). Utilizing Image Transforms and Diffusion Models for Generative Modeling of Short and Long Time Series. *Neural Information Processing Systems 2024*. https://doi.org/10.48550/arXiv.2410.19538

Nijkamp, E., Hill, M., Zhu, S.-C., & Wu, Y. N. (2019). Learning Non-Convergent Non-Persistent Short-Run MCMC Toward Energy-Based Model. *Neural Information Processing Systems*, *32*, 5232–5242.

Oppenheim, A. V., & Lim, J. S. (1981). The importance of phase in signals. *Proceedings of the IEEE*, *69*(5), 529–541. https://doi.org/10.1109/proc.1981.12022

Patel, Z., DeLoye, J., & Mathias, L. (2024). *Exploring Diffusion and Flow Matching Under Generator Matching*. https://doi.org/10.48550/arXiv.2412.11024

Pintilie, I., Manolache, A., & Brad, F. (2023). Time Series Anomaly Detection using Diffusion-based Models. *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, 570–578. https://doi.org/10.1109/icdmw60847.2023.00080

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1505.04597

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Ruiz-Cárcel, C., Cao, Y., Mba, D., Lao, L., & Samuel, R. T. (2015). Statistical process monitoring of a multiphase flow facility. *Control Engineering Practice*, *42*, 74–88. https://doi.org/10.1016/j.conengprac.2015.04.012

Sakai, S., & Hasegawa, T. (2025). *Reconstruction-Free Anomaly Detection with Diffusion Models via Direct Latent Likelihood Evaluation*. https://doi.org/10.48550/arXiv.2504.05662

Shiva, K., Etikani, P., Bhaskar, V. V. S. R., Mittal, A., Dave, A., Thakkar, D., Kanchetti, D., & Munirathnam, R. (2024). Anomaly Detection in Sensor Data with Machine Learning: Predictive Maintenance for Industrial Systems. *Journal of Electrical Systems*, *20*(10s), 454–462. https://journal.esrgroups.org/jes/article/view/5137

Xiao, Z., Yan, Q., & Amit, Y. (2021). *EBMs Trained with Maximum Likelihood are Generator Models Trained with a Self-adverserial Loss*. https://doi.org/10.48550/arXiv.2102.11757

Yang, S., Chen, Y., Wang, L., Liu, S., & Chen, Y. (2023). *Denoising Diffusion Step-aware Models*. https://doi.org/10.48550/arxiv.2310.03337

Yi, T.-H., Huang, H., & Li, H.-N. (2017). Development of sensor validation methodologies for structural health monitoring: A comprehensive review. *Measurement*, *109*, 200–214. https://doi.org/10.1016/j.measurement.2017.05.064

Yoon, S., Jin, Y.-U., Noh, Y.-K., & Park, F. C. (2023). Energy-based models for anomaly detection: a manifold diffusion recovery approach. *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 49445–49466.

Zamanzadeh Darban, Z., Webb, G. I., Pan, S., Aggarwal, C., & Salehi, M. (2024). Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, *57*(1), 1–42. https://doi.org/10.1145/3691338

Zhai, S., Cheng, Y., Lu, W., & Zhang, Z. (2016). Deep Structured Energy Based Models for Anomaly Detection. *In Proceedings of the 33rd International Conference on International Conference on Machine Learning*, *48*, 1100–1109. https://doi.org/10.48550/arxiv.1605.07717

Zhao, H., Zuo, S., Hou, M., Liu, W., Yu, L., Yang, X., & Deng, W. (2018). A Novel Adaptive Signal Processing Method Based on Enhanced Empirical Wavelet Transform Technology. *Sensors*, *18*(10), 3323. https://doi.org/10.3390/s18103323

Part 2: Research Paper

ScaloVit-EBM: Localized Energy-Based Anomaly Detection on Time–Frequency Scalograms

Shosuke Asano
School of Information Technology
Monash University Malaysia
Selangor, Malaysia
sasa0005@student.monash.edu

Abstract

Unsupervised anomaly detection in industrial time series is challenging due to non-stationarity, cross-channel interactions, and scarce anomaly labels. We present ScaloVit-EBM, a localized energy-based model that operates on multivariate sensor streams transformed into multichannel Continuous Wavelet Transform (CWT) scalograms. A U-Net encoder with a Vision Transformer produces per-patch energy scores, yielding interpretable time–frequency energy maps. The model is trained on normal-only data with Energy Matching, and the inference protocol uses overlapping chunking with max aggregation and a percentile threshold calibrated on normal validation data, decoupling scoring from decision-making. On the Cranfield Three-Phase Flow Facility benchmark, ScaloVit-EBM is competitive with or superior to LSTM EncDec-AD and PatchTrAD across multiple fault cases and excels on high-frequency slugging where localized time–frequency structure is critical. Ablations substantiate three practical choices: detrending improves class separability; max aggregation avoids diluting localized deviations across overlaps; and chunk width trades off long-range context against sensitivity to short-lived events. Additional analyses show that localized, per-patch scoring is crucial for spatially complex faults, while adding a contrastive-divergence term can raise F1 at the cost of ROC-AUC and compute. ScaloVit-EBM couples interpretable, localized energy scoring with a simple, deployment-minded decision rule, offering a robust foundation for real-world industrial anomaly detection.

1 Introduction

Unsupervised time-series anomaly detection (TSAD) underpins safety and reliability in industrial monitoring, where faults must be identified from multivariate sensor streams without labeled anomalies. Real-world signals are challenging: they are high-dimensional, non-stationary, and exhibit long-range dependencies and cross-channel interactions, all while labeled failures are scarce or unavailable [Liu et al., 2024, Boniol et al., 2024]. Contemporary TSAD approaches span three main families: reconstruction-based methods that flag large reconstruction errors [Malhotra et al., 2016, Audibert et al., 2020, Tuli et al., 2022]; forecasting-based methods that compare predictions to observations [Munir et al., 2019, Su et al., 2019]; and distribution-based methods that model normal data densities, such as normalizing flows [Rezende and Mohamed, 2015, Kingma and Dhariwal, 2018] and energy-based models (EBMs) [Du and Mordatch, 2019, Yoon et al., 2023] (see also Pang et al. [2021]). Despite progress, there remains a gap in jointly handling: (i) non-stationary, transient signatures; (ii) localized fault patterns that do not dominate all channels; and (iii) practical thresholding for deployment.

A central observation in this work is that representation matters as much as architecture. Purely time-domain models can miss evolving spectral content, whereas frequency-only analysis loses temporal localization [Oppenheim and Lim, 1981]. Time–frequency transforms address this trade-off. In particular, the Continuous Wavelet Transform (CWT) provides adaptive, multi-resolution analysis that is well-suited for non-stationary industrial signals with transient, multi-scale structure [Torrence and Compo, 1998, Mallat and Peyre, 2009, Rhif et al., 2019]. Casting multivariate signals into CWT magnitude scalograms yields a multichannel "image" that enables modern vision backbones to operate on localized time–frequency patterns.

We introduce ScaloVit-EBM, a localized energy-based model for multivariate industrial TSAD. The pipeline converts each sensor channel into a CWT magnitude scalogram and stacks features into a multichannel image. A U-Net encoder extracts hierarchical features, which are partitioned into non-overlapping patches and embedded into a Transformer. Rather than global pooling, a lightweight linear head produces a per-patch energy, yielding interpretable energy maps that highlight localized deviations in both time and frequency. For training on normal-only data, we adopt the Energy Matching objective of Balcerak et al. [2025], which instantiates Flow Matching [Lipman et al., 2023] for EBMs, and specialize it to our localized setting (multichannel CWT inputs and per-patch energies) while keeping the loss formulation unchanged. For ablation, we additionally test a variant that augments the objective with Contrastive Divergence (CD) while keeping the main model FM-only. At inference, overlapping chunking with max aggregation emphasizes the strongest local evidence, and a percentile threshold estimated from normal validation data provides a simple decision rule.

We evaluate on the Cranfield Three-Phase Flow Facility dataset [Ruiz-Cárcel et al., 2015], a realistic benchmark featuring multiple fault types with evolving dynamics. Against two representative baselines—LSTM EncDec-AD [Malhotra et al., 2016] and PatchTrAD [Vilhes et al., 2025]—ScaloVit-EBM achieves competitive or leading F1 and ROC-AUC on several fault cases and excels on high-frequency slugging, where localized time–frequency structure is critical. Ablations confirm three practical findings: (i) detrending improves class separability by suppressing low-frequency drift; (ii) max aggregation emphasizes the strongest local evidence and avoids dilution across overlaps; and (iii) chunk width trades off long-range context against sensitivity to short-lived events.

Our contributions are as follows:

- We propose ScaloVit-EBM, a patch-wise energy-based model operating on multichannel CWT scalograms, producing localized energy maps for interpretable anomaly localization in multivariate sensor data.
- We adopt Energy Matching [Balcerak et al., 2025] for normal-only industrial TSAD—retaining its objective while specializing the architecture and inference to localized per-patch energies—and include a CD-augmented ablation for comparison.
- We design a simple, deployment-friendly inference protocol—overlapping chunking, max aggregation, and percentile thresholding from normal validation—that separates scoring from decision-making.
- On the Cranfield MFP dataset, we demonstrate competitive results against LSTM EncDec-AD and PatchTrAD, with strong performance on complex, high-frequency faults; ablative and qualitative analyses clarify when representation, locality, and context length matter most.

The remainder of the paper is organized as follows. Section 2 reviews related work and background. Section 3 presents our methodology, including preprocessing, localized energy modeling, training, and inference. Section 4 reports experiments and results on the Cranfield MFP dataset, including baselines, ablations, and qualitative analyses. Section 5 discusses implications, limitations, and future work. Section 6 concludes.

2 Related work/Background

We position ScaloVit-EBM at the intersection of three complementary areas. First, we survey unsupervised time-series anomaly detection (TSAD), the primary field for this study; second, we review signal-to-image transformations that convert 1D temporal data into rich 2D representations; and third, we discuss Energy-Based Models (EBMs) and their suitability for modeling normal behavior and detecting out-of-distribution samples. By synthesizing these three pillars, we motivate our approach to detecting complex anomalies in multivariate sensor data.

2.1 Unsupervised Time-Series Anomaly Detection (TSAD)

Unsupervised Time-Series Anomaly Detection (TSAD) is a critical task across diverse applications, focusing on identifying unusual patterns in sequential data without requiring labeled anomaly examples. The inherent complexities of time-series, such as high dimensionality, temporal dependencies, and non-stationarity, pose significant challenges for effective anomaly detection [Liu et al., 2024, Boniol et al., 2024]. Early approaches to TSAD primarily leveraged classical and statistical methods such as One-Class Support Vector Machines (OC-SVM) [Schölkopf et al., 1999], Support Vector Data Description (SVDD) [Tax and Duin, 2004], and Isolation Forests [Liu et al., 2008] for outlier detection. However, these methods often struggle with high-dimensional, non-linear, and non-stationary time series, and may overlook complex temporal dependencies, thereby motivating the shift towards deep learning alternatives [Etikani et al., 2024].

The emergence of deep learning has revolutionized TSAD, providing powerful tools to model the intricate dynamics of time-series data. Deep learning-based TSAD methods can be broadly categorized by their operational mechanisms:

Reconstruction-Based Models. This prominent paradigm trains models to reconstruct normal time-series sequences, flagging significant reconstruction errors as anomalies. Early work in anomaly detection includes the LSTM-based Encoder-Decoder for Anomaly Detection (EncDec-AD) [Malhotra et al., 2016]. This method employs a single-layer LSTM in both its encoder and decoder to learn normal time-series patterns. For multivariate time-series, it represents each data point as an m-dimensional vector, which the model then learns to reconstruct. Advancements like USAD [Audibert et al., 2020] introduced adversarially-trained autoencoders for enhanced stability. More recently, Transformer-based architectures have gained traction for their ability to capture long-range dependencies; examples include PatchTrAD [Vilhes et al., 2025], which uses a patch-based Transformer encoder, and TranAD [Tuli et al., 2022], an adversarial Transformer encoder-decoder, both leveraging reconstruction for robust anomaly detection.

Forecasting-Based Models. This category identifies anomalies by detecting substantial deviations between predicted future time-series values and actual observations. DeepAnT [Munir et al., 2019] employs a CNN for next-step forecasting, while OmniAnomaly [Su et al., 2019] utilizes a GRU and VAE to learn a probabilistic representation and detect anomalies based on reconstruction likelihood. These predictive models implicitly capture sequence patterns, offering a complementary approach to reconstruction-based methods.

Probabilistic and Distribution-Based Models. This class explicitly models the underlying probability distribution of normal data. The Deep Autoencoding Gaussian Mixture Model (DAGMM) [Zong et al., 2018] combines an autoencoder with a GMM to estimate data density, flagging low-density regions as anomalous. Other density estimation techniques, such as Normalizing Flows [Rezende and Mohamed, 2015, Dinh et al., 2017, Kingma and Dhariwal, 2018], have also been explored for novelty detection, as they provide explicit likelihoods. Within this paradigm, Energy-Based Models (EBMs) offer a flexible framework by learning an energy function that assigns low energy to in-distribution (normal) data and high energy to out-of-distribution (anomalous) data, making them inherently well-suited for anomaly detection tasks [Nijkamp et al., 2019, Yoon et al., 2023]. These methods typically learn an explicit likelihood or an implicit mapping, offering a distinct approach compared to reconstruction or forecasting-based models [Pang et al., 2021].

2.2 Signal-to-Image Transformations for Time-Series Data

Given the complex and non-stationary temporal dynamics often present in sensor signals, their raw time-domain representation may not fully capture all relevant information [Zhao et al., 2018]. To extract more salient features and facilitate the application of powerful computer vision models, signal transformation techniques are frequently employed to convert one-dimensional time series into two-dimensional image-like representations [Naiman et al., 2024, Su et al., 2025]. Traditional time-domain analysis directly captures temporal patterns but can miss underlying periodicities. Conversely, frequency-domain analysis, such as the Fourier Transform [Oppenheim and Lim, 1981], reveals global spectral content but assumes stationarity and lacks temporal localization [Zhang et al., 2025]. This inherent trade-off between time and frequency resolution is a critical consideration when analyzing non-stationary signals.

To address these limitations, time-frequency representations offer a localized view of spectral content over time. The Short-Time Fourier Transform (STFT) [Oppenheim and Lim, 1981] applies a sliding window to segments of the signal, performing a Fourier Transform on each to generate a spectrogram. This provides insights into transient events and evolving spectral patterns. However, STFT is constrained by a fixed window size, which imposes a fundamental time-frequency resolution limit that can obscure fine temporal details [Zhang et al., 2025]. In contrast, the Continuous Wavelet Transform (CWT) utilizes scalable wavelet functions, enabling adaptive resolution across different frequencies [Mallat and Peyre, 2009]. CWT provides a continuous time-frequency representation that excels at decomposing non-stationary time series into their time-frequency components [Rhif et al., 2019], making it particularly well-suited for signals with evolving frequency content and transient events. While CWT can be more computationally intensive and may produce redundant data compared to STFT, its multi-resolution capability offers superior performance for capturing transient or non-stationary phenomena [Rhif et al., 2019].

Beyond time-frequency methods, other techniques transform time series into images to leverage the strengths of convolutional and vision-based models. Gramian Angular Fields (GAF) and Markov Transition Fields (MTF) [Wang and Oates, 2015] encode time series into 2D images by representing temporal correlations or transition probabilities, respectively. Recurrence Plots [MARWAN et al., 2007] visualize the recurrence of states in a phase space, revealing hidden patterns that can then be analyzed as images. These image-encoding techniques enable the application of advanced computer vision architectures to time-series data, effectively translating temporal dynamics into spatial patterns.

2.3 Energy-Based Models (EBMs) for Anomaly Detection

Energy-Based Models (EBMs) present a robust and adaptable framework for learning intricate, high-dimensional data distributions, rendering them particularly suitable for anomaly detection tasks [Du and Mordatch, 2019, Nijkamp et al., 2019, Yoon et al., 2023]. The fundamental concept behind EBMs is the definition of an energy function, $E(x)$, typically implemented via a neural network. This function assigns low energy values to data points considered normal (in-distribution) and higher energy values to those deemed anomalous (out-of-distribution) [Du and Mordatch, 2019, Yoon et al., 2023]. Consequently, anomalies are often found in regions of the data manifold associated with elevated energy scores. Formally, EBMs implicitly define a probability distribution over data x through a Boltzmann distribution: $P(x) = \frac{e^{-E(x)}}{Z}$, where $Z = \int e^{-E(x)}dx$ is the intractable partition function.

For anomaly detection, the core inference step involves calculating $E(x)$ for a given test input. An input is classified as anomalous if its computed energy surpasses a predetermined threshold [Yoon et al., 2023]. This direct method of energy evaluation for anomaly scoring is a standard practice. EBMs also offer a degree of interpretability by directly modeling energy landscapes that hold semantic meaning [Carbone, 2025]. Furthermore, their unsupervised learning paradigm naturally aligns with real-world anomaly detection scenarios where labeled anomalous data are scarce or unavailable.

A significant challenge in EBM training stems from the intractability of the partition function Z, which makes maximum likelihood training difficult [Xiao et al., 2021]. To circumvent this, various approximate training methods have been developed:

- **Contrastive Divergence (CD)** [Hinton, 2012]: A classical algorithm that approximates the likelihood gradient using a short Markov Chain Monte Carlo (MCMC) chain. While making EBMs tractable, CD can suffer from slow mixing and biased updates.

- **Score Matching** [Zhai et al., 2016]: This approach avoids the partition function by minimizing the difference between the score functions of the model and the empirical data distribution.

- **Flow Matching / Energy Matching** [Balcerak et al., 2025, Loo et al., 2025, Lipman et al., 2023]: Recent advancements aim to unify EBMs with flow-based sampling. These frameworks learn a single scalar energy field such that samples flow via optimal transport towards the data manifold, concentrating energy into a Boltzmann distribution. This approach simplifies training by removing the need for auxiliary diffusion timesteps and offers improved sample fidelity.

EBMs offer practical advantages for anomaly detection. Unlike some likelihood-based models such as VAEs and normalizing flows, which may assign high likelihoods to outliers in certain regimes, EBMs trained with iterative negative sampling can sharpen decision boundaries around the normal manifold and achieve strong out-of-distribution detection [Yoon et al., 2023]. Moreover, because EBMs model unnormalized densities directly, they can be advantageous for identifying low-density anomalies and can be less susceptible to certain failure modes observed elsewhere.

3 Methodology

The proposed methodology for localized anomaly detection in multivariate time-series sensor data comprises three main stages: (1) signal-to-image transformation, where each 1D feature signal is converted into a 2D time–frequency representation and concatenated into a multichannel tensor; (2) a patch-based energy model, which assigns an energy score to localized regions of the resulting multichannel image; and (3) an anomaly scoring and localization process, which aggregates patch-level energies to identify anomalous events over time.

3.1 Signal Preprocessing and Time–Frequency Representation

Let the raw multivariate soft-sensor time-series data be denoted as

$$S = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_L\}, \quad \mathbf{s}_t \in \mathbb{R}^F$$

where F is the number of sensor features and L the total sequence length.

Detrending. To focus on short-term fluctuations and suppress slow baseline drift, each feature channel is individually detrended using a moving average filter with window size W_{MA} [Mitov, 1998]. For each feature $f \in \{1, \dots, F\}$, the residual signal is:

$$s^{(f)}_{\text{resid}}(t) = s^{(f)}(t) - \frac{1}{W_{\text{MA}}} \sum_{i=-(W_{\text{MA}}-1)/2}^{(W_{\text{MA}}-1)/2} s^{(f)}(t+i)$$

This ensures that low-frequency variations do not dominate the subsequent time–frequency decomposition and maintains consistency across features with different dynamic ranges. Each detrended feature is then normalized via min–max scaling to [-1, 1].

Continuous Wavelet Transform (CWT). Each normalized feature is independently transformed into a 2D time–frequency representation using the Continuous Wavelet Transform (CWT) [Torrence and Compo, 1998]:

$$C^{(f)}(a,b) = \int_{-\infty}^{\infty} s^{(f)}(t) \frac{1}{\sqrt{a}} \psi^* \left(\frac{t-b}{a} \right) dt$$

where a denotes the scale (inversely related to frequency), b the translation (time), and ψ^* the complex conjugate of the Morlet mother wavelet. The magnitude scalogram is then defined as $M^{(f)}(a,b) = |C^{(f)}(a,b)|$.

All feature-wise scalograms are stacked along the channel dimension to form a multichannel time–frequency tensor:

$$M = \text{Concat}(M^{(1)}, M^{(2)}, \dots, M^{(F)}) \in \mathbb{R}^{F \times H \times W}$$

where H and W denote the number of scales and time samples, respectively. This representation preserves inter-feature relationships while providing fine-grained temporal and frequency localization for non-stationary sensor behaviors.

Overall Architecture Workflow. The overall model workflow is illustrated in Figure 1. Starting from the raw multivariate sensor signal, the data are sequentially detrended and normalized to suppress low-frequency drift. Each feature is then converted into a time–frequency scalogram using the Continuous Wavelet Transform (CWT), and the resulting multichannel tensor $M \in \mathbb{R}^{F \times H \times W}$ serves as the model input.

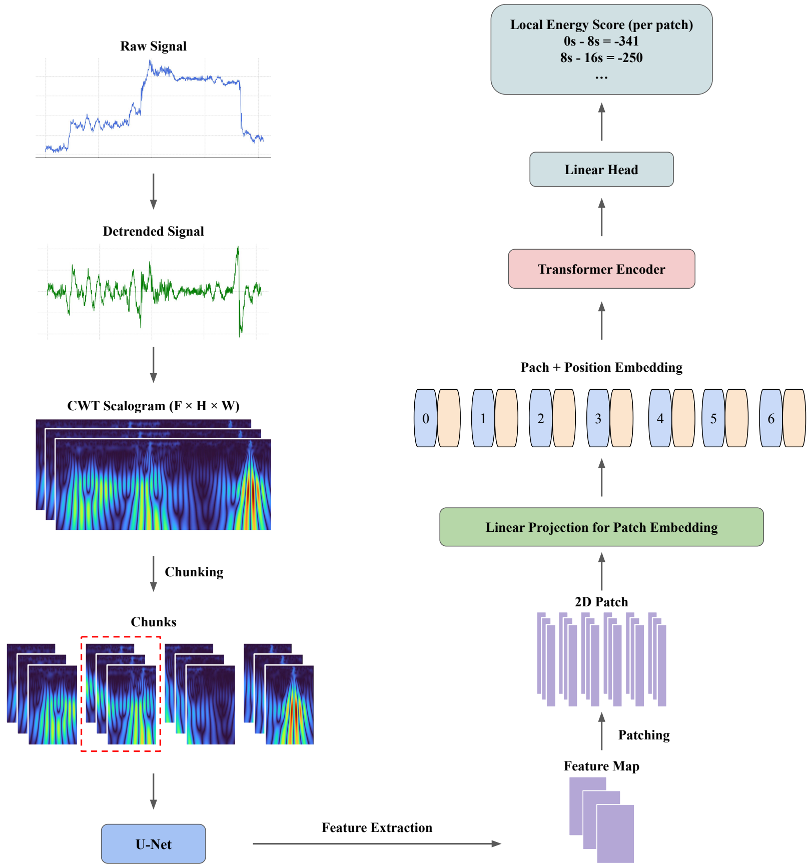


Figure 1: Overview of the proposed model architecture. The workflow begins with raw multivariate sensor data, followed by detrending and CWT-based time–frequency transformation. Each chunk of the resulting multichannel scalogram is processed through a U-Net feature extractor, patch division, and a Transformer encoder. The model outputs localized energy scores per patch, forming interpretable energy maps that enable precise anomaly localization.

During processing, the input scalogram is divided into overlapping temporal chunks, each passed through the following sequence:

$$\begin{aligned} &\text{Chunk} \rightarrow \text{U-Net (Feature Extraction)} \rightarrow \text{Patch Division} \\ &\quad \rightarrow \text{Linear Patch Embedding + Positional Encoding} \\ &\quad \rightarrow \text{Transformer Encoder} \rightarrow \text{Linear Head} \end{aligned}$$

This pipeline produces a patch-level energy score for each localized region within the chunk. The scores are then reconstructed into a continuous energy map that aligns with the original time–frequency coordinates. For clarity, the complete end-to-end architecture — from signal preprocessing to local energy computation — is summarized in Figure 1.

3.2 Localized Energy-Based Model

The core of the method is an Energy-Based Model (EBM) that learns the joint distribution of normal, multivariate sensor behavior. An EBM defines a probability distribution via an energy function $E_\theta(x)$, parameterized by θ:

$$p_\theta(x) = \frac{e^{-E_\theta(x)}}{Z_\theta}, \quad Z_\theta = \int e^{-E_\theta(x)} dx$$

Normal samples correspond to low-energy regions of the learned distribution, while anomalous patterns yield high energy values.

The model architecture is adapted from the Energy Matching framework proposed by Balcerak et al. [2025] and modified for localized anomaly detection in multichannel time–frequency representations. Specifically, we preserve the original energy formulation and training objective, while adjusting the encoder to process multi-feature CWT patches and to capture localized dependencies across both time and sensor dimensions.

Image Patching. The multichannel scalogram $M \in \mathbb{R}^{F\times H\times W}$ is divided into a grid of N non-overlapping patches [Dosovitskiy et al., 2021]:

$$\{p_1, p_2, \ldots, p_N\}, \quad p_i \in \mathbb{R}^{F\times P_H \times P_W}, \quad N = (H/P_H) \times (W/P_W)$$

This patch-based design captures localized inter-feature and temporal–spectral dependencies, enabling the model to detect subtle, spatially confined anomalies that may not affect the entire feature set simultaneously.

Architecture and Local Energy Computation. The input multichannel scalogram is first processed by a U-Net backbone to extract hierarchical feature representations that capture spatial–temporal context across sensor channels. The resulting feature map is then divided into non-overlapping patches, each of which is linearly embedded and augmented with a positional encoding before being passed through a Transformer encoder T_θ.

In the original Energy Matching [Balcerak et al., 2025], the output tokens are mean-pooled to produce a single global energy:

$$\mathbf{z}_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N} T_\theta(p_i), \quad E_{\text{global}} = \text{Linear}(\mathbf{z}_{\text{avg}})$$

However, global pooling suppresses local variations critical for anomaly localization. To retain spatial–temporal information, we remove mean-pooling and compute a local energy score per patch:

$$E_{\text{local}}(p_i) = \text{Linear}(T_\theta(p_i))$$

This yields a set of patch-level energy values $\{E_{\text{local}}(p_1), \ldots, E_{\text{local}}(p_N)\}$, providing interpretable energy maps that highlight localized deviations across both feature and time–frequency dimensions.

3.3 Model Training via Energy Matching

We train the EBM on normal-only data using Conditional Flow Matching (CFM). To assess the impact of additional negative-sample pressure, we also evaluate a variant that augments the objective with Contrastive Divergence (CD). The main model reported in our experiments uses CFM only; the `w/ CD Loss` variant adds CD for an ablation comparison.

Flow Matching Loss. CFM aligns the model's velocity field $v_\theta(x,t) = -\nabla_x E_\theta(x,t)$ with a target velocity u_t derived from an optimal transport path between noise and data distributions [Lipman et al., 2023]:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,p(x_t|x_1)}\left[\|v_\theta(x_t,t) - u_t(x_t|x_1)\|^2\right]$$

This regularizes the model to form a smooth energy landscape around normal samples, improving convergence stability.

Contrastive Divergence Loss. The CD loss pushes the model to assign low energy to real (normal) data and high energy to negative samples generated via Gibbs sampling:

$$\mathcal{L}_{\text{CD}} = \mathbb{E}_{p_{\text{data}}(x_{\text{pos}})}[E_\theta(x_{\text{pos}})] - \mathbb{E}_{p\theta(x_{\text{neg}})}[E_\theta(x_{\text{neg}})]$$

In our main model we set $\lambda_{\text{CD}} = 0$ (FM-only); the ablation variant `w/ CD Loss` uses $\lambda_{\text{CD}} > 0$ while keeping all other settings identical.

The total objective integrates both terms with gradient regularization:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{flow}}\mathcal{L}_{\text{flow}} + \lambda_{\text{CD}}\mathcal{L}_{\text{CD}} + \lambda_{\text{reg}}\|\nabla_x E_\theta(x)\|^2$$

Unless otherwise stated, we use FM-only in all main experiments (i.e., $\lambda_{\text{CD}} = 0$); the `w/ CD Loss` ablation uses $\lambda_{\text{CD}} > 0$. This training objective encourages the model to learn a robust energy manifold that tightly envelops the distribution of normal multivariate sensor behavior, ensuring that any deviation from this manifold results in a high energy score, indicative of an anomaly.

3.4 Anomaly Detection and Localization

During inference, a long multivariate signal is processed using a sliding-window strategy.

Chunking and Scoring. The full multichannel scalogram is divided into overlapping chunks of width W_{chunk} and stride S_{chunk}. Each chunk is passed through the trained EBM to compute patch-level energy scores. Overlapping windows ensure that anomalies near chunk boundaries are not missed and that energy continuity is preserved across time.

Score Aggregation. Patch-level energies from overlapping chunks are reassembled into a continuous energy tensor aligned with the time–frequency coordinates:

$$E_{\text{timeline}}(t) = \max_{\text{chunks containing } t} E_{\text{local}}(t)$$

The max aggregation emphasizes the strongest anomaly evidence across overlapping regions, enabling high sensitivity to localized disturbances.

Thresholding. An anomaly threshold τ is determined from a validation set containing only normal data (e.g., the 99.5th percentile of energy values). A time step t is classified as anomalous if its aggregated energy exceeds the threshold:

$$\text{Anomaly}(t) = \begin{cases} 1, & E_{\text{timeline}}(t) > \tau \\ 0, & \text{otherwise} \end{cases}$$

This approach provides a comprehensive anomaly timeline, enabling fine-grained localization of anomalous events across both time and feature dimensions in the original multivariate signal.

4 Experiments and Results

4.1 Dataset Description

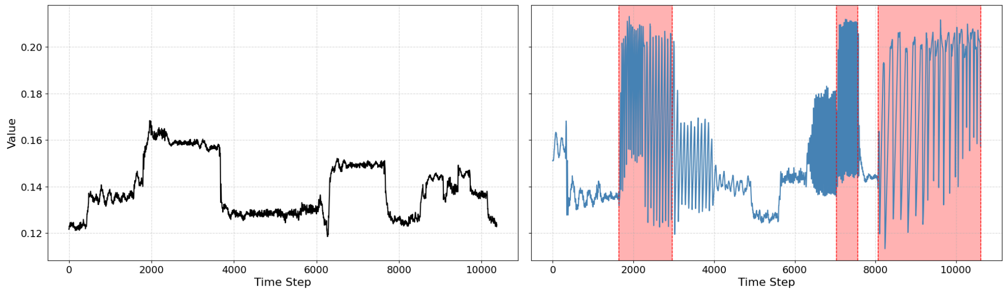
This study utilizes the *Three-Phase Flow Facility* dataset from Cranfield University [Ruiz-Cárcel et al., 2015], which records the dynamics of a controlled two-phase (air–water) flow system. The dataset was collected from a pressurized experimental rig equipped with pipelines of varying diameters, separators, and tanks, all managed by a Delta V SCADA system. Measurements were recorded at a sampling frequency of 1 Hz.

A total of 24 process variables are provided, including flow rates, pressures, densities, and liquid levels. In this study, 23 soft-sensor variables were used, excluding PT417 (pressure in the 2" line) since it is only relevant to Fault 6. The resulting multivariate time-series data are well suited for evaluating data-driven fault detection algorithms.

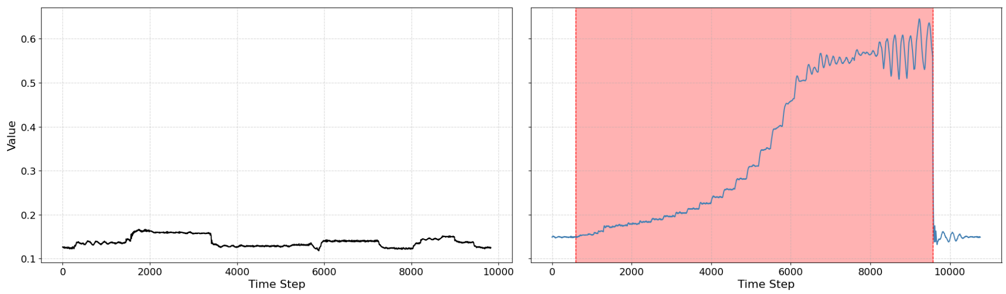
Training Data (Normal Operation). The training set consists of three normal-condition acquisitions (T1–T3), each representing the system under varying air and water flow conditions. These datasets were collected across 20 combinations of air and water setpoints, capturing large and small transients in both increasing and decreasing directions. This setup ensures that the training data encompass a diverse range of normal operating dynamics.

Testing Data (Faulty Operation). The testing set comprises 14 datasets corresponding to five fault types, each introduced after a period of normal operation:

1. **Air line blockage** (Fault case 1.1–1.3),
2. **Water line blockage** (Fault case 2.1–2.3),
3. **Top separator input blockage** (Fault case 3.1–3.3),



(a) Normal (T1, Feature 1) vs. Slugging (Fault case 5.2), Feature 1



(b) Normal (T2, Feature 1) vs. Input Blockage (Fault case 3.3), Feature 1

Figure 2: Representative examples of sensor signals for Feature 1 (pressure). Each subplot compares a normal operation signal (left) with a faulty one (right). Red-shaded regions indicate the fault intervals. (a) Compares normal operation (T1) with slugging conditions (Fault case 5.2). (b) Compares normal operation (T2) with a top separator input blockage (Fault case 3.3).

4. **Open direct bypass simulating a leak** (Fault case 4.1–4.3),
5. **Slugging conditions** (Fault case 5.1–5.2).

Each fault scenario was seeded gradually to allow observation of the fault progression over time. The sixth fault case, which involves pressurization of an isolated 2" line, was excluded from this study because it relies on the PT417 variable, which is not included among the 23 selected sensors.

Each dataset includes clear timestamps marking fault onset and recovery, enabling quantitative evaluation of fault detection performance. In total, the dataset provides three normal-condition sequences for training and fourteen fault-condition sequences for testing.

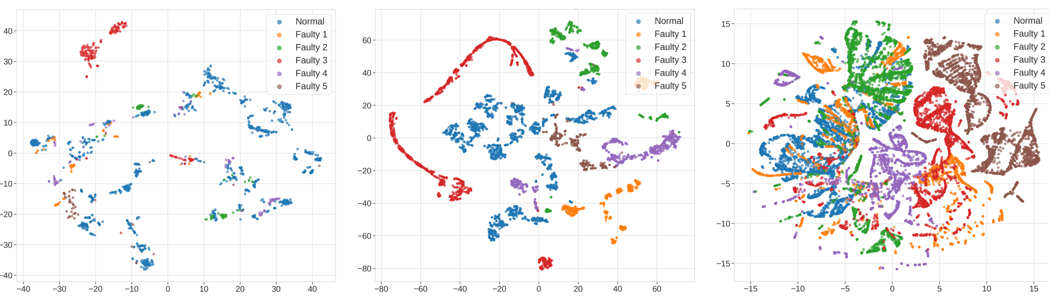
4.2 Quantitative Comparison with Baselines

To rigorously assess the effectiveness of the proposed model, we compare it against two representative baselines: the LSTM-based Encoder-Decoder for Anomaly Detection (LSTM EncDec-AD) by [Malhotra et al., 2016] and the Patch-based Transformer architecture (PatchTrAD) proposed by [Vilhes et al., 2025]. These models embody two complementary paradigms in unsupervised time-series anomaly detection—temporal sequence reconstruction through recurrent networks and patch-wise reconstruction-based modeling using Transformer encoders. All models are trained and evaluated on the same preprocessed Multiphase Flow Process (MFP) dataset under identical experimental settings. Performance is assessed using two principal metrics: F1 Score and ROC-AUC. The F1 Score quantifies the balance between precision and recall under a fixed threshold, while ROC-AUC provides a threshold-independent measure of discriminative capability. Together, these metrics capture complementary aspects of anomaly detection performance. Table 1 summarizes the results across all fault cases, with bold values indicating the best-performing model for each metric and scenario.

Overall, the comparison reveals that no single model consistently dominates across all scenarios. The LSTM EncDec-AD performs particularly well in FaultyCase3 and FaultyCase4, achieving near-perfect F1 and ROC-AUC scores in some subsets, reflecting its strength in capturing longer-term

Table 1: Performance comparison of baseline and proposed models across all fault cases in the MFP dataset, focusing on F1 Score and ROC-AUC. Bold values indicate the best performance for each metric.

Fault Case	**LSTM EncDec-AD [Malhotra et al., 2016]**		**PatchTrAD [Vilhes et al., 2025]**		**Proposed (Ours)**	
	F1	ROC-AUC	F1	ROC-AUC	F1	ROC-AUC
FaultyCase1_Set1_1	0.64	0.61	0.49	0.68	**0.78**	**0.74**
FaultyCase1_Set1_2	0.56	0.61	0.50	0.54	**0.86**	**0.65**
FaultyCase1_Set1_3	0.55	0.42	0.37	0.56	**0.84**	**0.65**
FaultyCase2_Set2_1	0.33	0.48	0.00	**0.52**	**0.45**	0.37
FaultyCase2_Set2_2	**0.26**	0.18	0.00	0.36	0.03	**0.49**
FaultyCase2_Set2_3	0.24	0.17	0.01	0.36	**0.66**	**0.40**
FaultyCase3_Set3_1	**0.98**	**0.99**	**0.98**	0.94	0.89	0.81
FaultyCase3_Set3_2	**0.95**	**0.74**	0.69	0.54	0.74	0.45
FaultyCase3_Set3_3	0.96	**0.99**	**0.98**	0.92	0.91	0.37
FaultyCase4_Set4_1	**0.89**	**0.80**	0.59	0.77	0.85	0.71
FaultyCase4_Set4_2	0.47	**0.61**	0.26	0.36	**0.81**	**0.61**
FaultyCase4_Set4_3	0.70	0.41	0.36	0.39	**0.90**	**0.67**
FaultyCase5_Set5_1	**0.69**	0.59	0.59	0.66	0.57	**0.94**
FaultyCase5_Set5_2	0.84	**0.97**	**0.87**	0.96	0.59	0.96
Overall Average	**0.77**	**0.77**	0.67	0.73	0.76	0.60



(a) LSTM EncDec-AD (b) PatchTrAD (c) Proposed Model

Figure 3: Comparative t-SNE visualization of latent feature spaces from (a) the LSTM EncDec-AD, (b) PatchTrAD, and (c) the proposed model. Each point is a time-series segment colored by its true label (Normal or Fault types 1-5). The visualization highlights the complementary strengths of the models: while the baseline LSTM EncDec-AD (a) effectively isolates FaultyCase 3, the proposed model (c) uniquely excels at separating the complex patterns of FaultyCase 5 from normal behavior, even though some overlap persists in other classes. This demonstrates our model's capability to learn discriminative features for specific, challenging anomaly signatures.

temporal dependencies in quasi-periodic faults. In contrast, its performance notably declines in FaultyCase2, suggesting difficulty handling less predictable or irregular signals. The PatchTrAD model exhibits more stable results across fault conditions, often achieving moderate to high ROC-AUC values, particularly in FaultyCase3 and FaultyCase5, where localized representations help discriminate subtle variations in temporal patterns. However, it occasionally underperforms in F1 Score when threshold calibration becomes challenging due to overlapping normal–fault distributions. The proposed model achieves competitive or leading F1 and ROC-AUC values in several fault cases (e.g., FaultyCase1 and FaultyCase4), indicating its effectiveness in detecting deviations from the learned distribution of normal behavior. Nevertheless, its performance fluctuates in certain subsets such as FaultyCase3, suggesting that while the patch-level representations effectively capture local anomalies, they may be less robust in globally periodic patterns compared to recurrent baselines.

To further interpret these quantitative findings, we visualize the latent feature distributions using t-distributed Stochastic Neighbor Embedding (t-SNE) in Fig.3. Each point represents a segment of the time series, color-coded by its true label (Normal or Faulty 1–5). The LSTM EncDec-AD (Fig.3a) exhibits considerable overlap between normal and faulty representations, except for FaultyCase 3, which forms a distinct cluster—consistent with its strong numerical performance in Table 1. The PatchTrAD model (Fig.3b) demonstrates improved separation between normal and fault samples, particularly for FaultyCase 3, though partial overlap remains for FaultyCase 4 and FaultyCase 5. In contrast, the proposed model (Fig.3c) clearly distinguishes FaultyCase 5 from normal samples, aligning with its outstanding ROC-AUC score in that scenario. However, partial intermixing persists among FaultyCase 1–4 and normal representations, indicating potential room for refinement in feature disentanglement for overlapping fault types.

Collectively, these results reveal a clear architectural trade-off: the recurrent LSTM is superior for globally periodic faults, while the patch-based PatchTrAD is effective for localized events. Our proposed energy-based model demonstrates a powerful synthesis, achieving a balanced performance across diverse faults while uniquely excelling at complex, high-frequency anomalies like FaultyCase 5.

4.3 Ablation Study

To validate our core design choices, we conducted an extensive ablation study evaluating the impact of architectural and methodological variations on overall performance. We compare our full proposed model, **ScaloVit-EBM**, against four ablated variants:

- **w/o Detrending**: Removes the moving-average detrending step (Section 3.1) from the preprocessing pipeline.
- **Global-ViT (Image-Based Scoring)**: Replaces the per-patch scoring mechanism (Section 3.2) with a global score, testing the value of localized energy computation.
- **w/ CD Loss**: Trained with the Contrastive Divergence (CD) loss component (Section 3.3) to evaluate its contribution.
- **Mean Aggregation**: Replaces the `max` aggregation strategy (Section 3.4) with `mean` averaging for reconstructing scores.

Table 2: Detailed ablation study results across all fault cases. For each case, we report F1 Score and ROC-AUC. Bold values indicate the best-performing model variant for that metric in that row.

Fault Case	**ScaloVit-EBM (FM-only)**		**w/o Detrending**		**Global-ViT**		**w/ CD Loss**		**Mean Aggregation**	
	F1	ROC-AUC	F1	ROC-AUC	F1	ROC-AUC	F1	ROC-AUC	F1	ROC-AUC
FaultyCase1_Set1_1	**0.78**	**0.74**	0.75	0.60	0.77	0.71	0.77	0.70	0.63	0.53
FaultyCase1_Set1_2	**0.86**	**0.65**	0.83	0.59	**0.86**	0.62	**0.86**	**0.65**	0.71	0.51
FaultyCase1_Set1_3	0.84	**0.65**	0.82	0.59	0.84	**0.65**	**0.85**	0.64	0.68	0.54
FaultyCase2_Set2_1	0.45	0.37	0.00	0.04	0.45	**0.41**	**0.46**	0.38	0.16	0.27
FaultyCase2_Set2_2	0.03	0.49	0.57	0.36	0.61	**0.55**	**0.78**	0.45	0.01	0.38
FaultyCase2_Set2_3	0.66	0.40	**0.77**	0.24	0.73	**0.46**	0.75	0.35	0.06	0.27
FaultyCase3_Set3_1	**0.89**	0.81	**0.89**	0.64	**0.89**	**0.85**	**0.89**	0.67	**0.89**	0.46
FaultyCase3_Set3_2	0.74	**0.45**	0.69	0.29	0.67	0.42	**0.76**	0.43	0.62	0.38
FaultyCase3_Set3_3	**0.91**	**0.37**	**0.91**	0.34	**0.91**	0.27	**0.91**	0.34	**0.91**	0.17
FaultyCase4_Set4_1	**0.85**	**0.71**	**0.85**	0.61	**0.85**	0.43	**0.85**	0.52	0.65	0.53
FaultyCase4_Set4_2	**0.81**	0.61	**0.81**	0.54	**0.81**	**0.70**	**0.81**	0.62	**0.81**	0.16
FaultyCase4_Set4_3	**0.90**	**0.67**	**0.90**	0.26	**0.90**	**0.67**	**0.90**	0.40	**0.90**	0.26
FaultyCase5_Set5_1	**0.57**	**0.94**	**0.57**	**0.94**	**0.57**	0.65	**0.57**	0.92	**0.57**	0.90
FaultyCase5_Set5_2	0.59	**0.96**	0.59	**0.96**	0.60	0.92	0.59	0.95	**0.63**	0.95
Overall Average	0.76	**0.60**	0.76	0.57	0.77	0.59	**0.78**	0.59	0.69	0.55

Analysis of Architectural Components. The detailed results in Table 2 reveal several key insights. The clearest findings relate to the preprocessing and aggregation steps. The choice of aggregation strategy proves critical, as the Mean Aggregation model yields the lowest overall ROC-AUC score of all variants (0.55). A score this close to random chance (0.5) indicates that mean averaging dilutes localized, high-energy deviations, making the normal and faulty distributions nearly indistinguishable. This strongly validates our use of max aggregation to preserve these critical signatures. Similarly, removing the preprocessing step in the w/o Detrending model causes the overall ROC-AUC to fall from 0.60 to 0.57. The performance degradation is particularly severe in certain scenarios, such as `FaultyCase2_Set2_1` and `FaultyCase4_Set4_3`, where the ROC-AUC collapses to near-random levels (0.04 and 0.26, respectively). This suggests that without removing the natural baseline variations present in the raw signals, the model struggles to learn a discriminative energy function that can reliably separate anomalous patterns from normal operational fluctuations. Together, these findings confirm that both detrending and max aggregation are essential for learning a robust and discriminative energy function.

Value of Localized Scoring. The Global-ViT ablation, which replaces per-patch scoring with a single global energy score per chunk, directly tests our core architectural hypothesis. At first glance, its overall performance appears comparable to the full model, with a slightly higher F1 score (0.77) and a nearly identical ROC-AUC (0.59). However, this aggregate view masks a critical weakness. For FaultyCase5 (slugging), the performance of the Global-ViT model collapses, with its ROC-AUC score dropping from 0.94 to 0.65 in one instance. This demonstrates that while a global energy score is sufficient for many fault types, the localized, per-patch scoring mechanism is indispensable for correctly identifying spatially complex anomalies like slugging, thus validating our proposed architecture.

Impact of Loss Function. The comparison between the top-performing variants reveals a crucial design trade-off. Our main proposed model, ScaloVit-EBM (FM-only), achieves the highest overall ROC-AUC (0.60), indicating the best intrinsic class separability. In an ablation study, we introduced a computationally intensive Contrastive Divergence (CD) loss (the w/ CD Loss model). Interestingly, this variant achieved the highest overall F1 score (0.78). This presents a trade-off between performance metrics and computational cost. While adding the CD loss can create a decision boundary that is highly effective for a single-threshold detector (leading to a higher aggregate F1), it does so at the cost of a slight decrease in overall class separability (lower ROC-AUC) and a significant increase in training complexity. The strong performance of our proposed model without this extra loss validates our design choice for a more efficient architecture that maintains superior class-separation capabilities.

Impact of Chunk Size. To evaluate the effect of temporal context, we experimented with different chunk widths for the input scalograms. Table 3 provides a detailed comparison of our model's performance using chunk widths of 2048 (default), 1024, and 512.

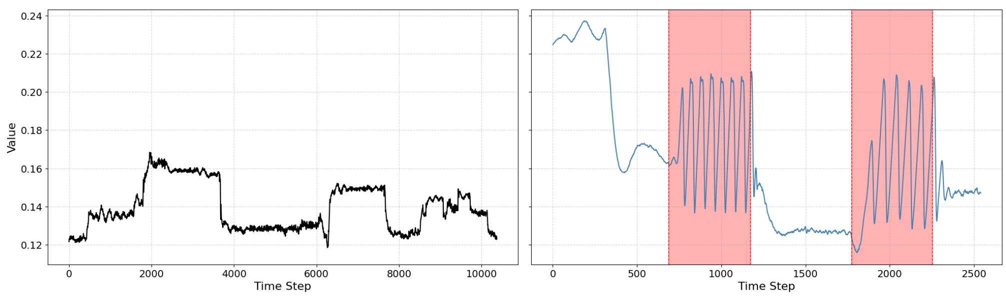
Table 3: Detailed performance comparison for different chunk widths across all fault cases. Bold values indicate the best-performing chunk width for that metric in that row.

Fault Case	**2048 (Default)**		**1024**		**512**	
	F1	ROC-AUC	F1	ROC-AUC	F1	ROC-AUC
FaultyCase1_Set1_1	**0.78**	**0.74**	0.54	0.73	0.52	0.43
FaultyCase1_Set1_2	**0.86**	**0.65**	0.63	0.64	0.57	0.36
FaultyCase1_Set1_3	**0.84**	**0.65**	0.61	0.51	0.48	0.43
FaultyCase2_Set2_1	**0.45**	0.37	0.26	**0.53**	0.17	0.42
FaultyCase2_Set2_2	0.03	0.49	**0.54**	**0.80**	0.26	0.23
FaultyCase2_Set2_3	**0.66**	0.40	0.02	**0.85**	0.27	0.25
FaultyCase3_Set3_1	**0.89**	**0.81**	0.87	0.58	0.80	0.73
FaultyCase3_Set3_2	**0.74**	**0.45**	0.53	0.39	0.51	0.41
FaultyCase3_Set3_3	**0.91**	0.37	0.84	0.33	0.76	**0.48**
FaultyCase4_Set4_1	**0.85**	**0.71**	0.79	0.60	0.69	0.36
FaultyCase4_Set4_2	**0.81**	**0.61**	0.78	0.60	0.68	0.51
FaultyCase4_Set4_3	**0.90**	**0.67**	0.84	0.51	0.74	0.28
FaultyCase5_Set5_1	0.57	0.94	0.57	0.90	0.57	**0.95**
FaultyCase5_Set5_2	0.59	0.96	0.70	0.96	**0.75**	**0.97**
Overall Average	**0.76**	0.60	0.69	**0.61**	0.63	0.56

Table 3, reveals a fundamental trade-off between capturing long-range dependencies and sensitivity to localized events. The largest chunk size, 2048, achieves the best overall F1 score (0.76), suggesting that a wider temporal context is crucial for modeling the slow-drifting dynamics and long-range patterns characteristic of normal operation. However, this larger window can dilute the signal of short-lived, abrupt anomalies. This is evidenced by the superior performance of smaller chunk sizes on specific fault types. For instance, the 1024-width chunk consistently excels across all subsets of FaultyCase2. Most notably, the 512-width chunk is the top performer on the high-frequency slugging anomaly in `FaultyCase5_Set5_2`, boosting the F1 score from 0.59 to 0.75. This occurs because the anomaly constitutes a larger, more potent portion of the input in a smaller window. These results demonstrate that while a larger chunk size is a robust default, the optimal choice is dependent on the temporal scale of the target anomaly, highlighting chunk size as a critical hyperparameter for practical application. These findings suggest practitioners should tune W_{chunk} to match the temporal scale of anticipated faults: larger windows favor slow dynamics, while smaller windows increase sensitivity to short-lived anomalies.

4.4 Qualitative Analysis of Anomaly Patterns

To gain deeper insight into the model's detection capabilities, we perform a qualitative analysis on two representative fault cases that highlight its differential performance: `FaultyCase5_Set5_1` (slugging), where the model excels, and `FaultyCase2_Set2_1` (water line blockage), where it struggles. Figure 4 contrasts the raw signal characteristics of these two faults, which helps explain their differing detection outcomes.



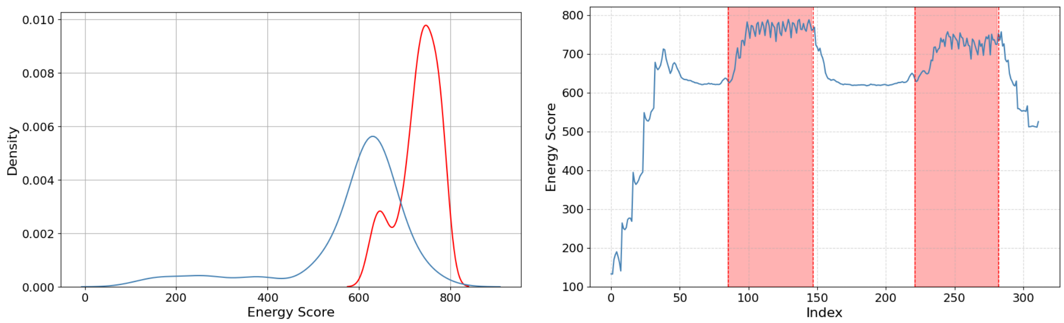
(a) Normal (T1, Feature 1) vs. Slugging (`FaultyCase5_Set5_1`, Feature 1)

(b) Normal (T2, Feature 1) vs. Water Line Blockage (`FaultyCase2_Set2_1`, Feature 1)

Figure 4: Representative examples of sensor signals for Feature 1 (pressure), selected to illustrate the model's differential performance. Each subplot compares a normal operation signal (left) with a faulty one (right). Red-shaded regions indicate the ground-truth fault interval. (a) Compares normal operation (T1) with chaotic signature of slugging (`FaultyCase5_Set5_1`), a fault type the model detects effectively. (b) Compares normal operation (T2) with gradual drift characteristic of a water line blockage (`FaultyCase2_Set2_1`), which poses a greater detection challenge.

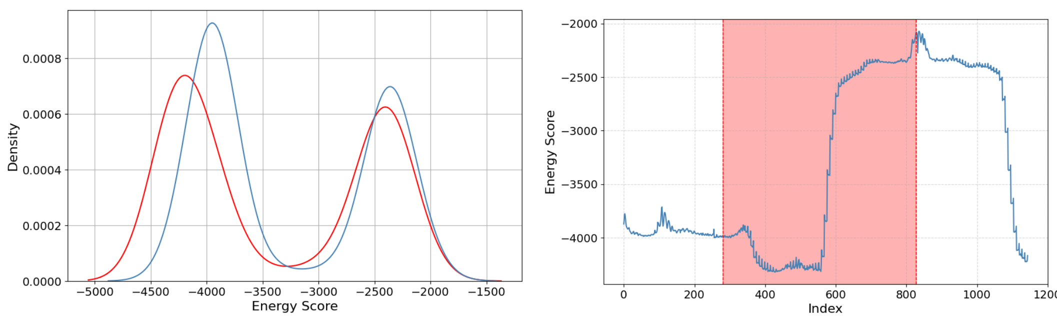
Analysis of High-Frequency Anomaly Detection (Slugging). Figure 5 exemplifies the model's strength in detecting the slugging conditions in `FaultyCase5_Set5_1`. The energy distribution (Figure 5a) shows a clear and wide separation between the low-energy manifold of normal data and the high-energy region of anomalous data. This clean separation corresponds to the high ROC-AUC score (0.94) reported in Table 1, indicating the model has learned a highly discriminative energy function. Consequently, the energy score over time (Figure 5b) is precise and decisive; it rises at the fault's onset and returns to the normal baseline immediately upon resolution, confirming the model's effectiveness in localizing transient, non-stationary anomalies.

Analysis of Slow-Drift Anomaly Detection (Blockage). In contrast, Figure 6 depicts the model's struggles with the gradual water line blockage in `FaultyCase2_Set2_1`. The energy distribution (Figure 6a) reveals limited separability for this fault type. Instead of assigning consistently higher energy to the anomaly, the model produces heavily overlapping distributions where the anomalous scores (red) are not clearly separated to the right of the normal scores (blue). This indicates that the learned energy function is not discriminative for this type of deviation, which directly explains the near-random ROC-AUC score (0.37) for this case (Table 1). The energy timeline (Figure 6b) further exposes this limitation; instead of a prompt response, the model's energy score remains low during the initial fault phase and rises only gradually. This delayed and weaker response degrades detection performance and highlights the model's difficulty with slowly evolving deviations.



(a) Energy distribution (`FaultyCase5_Set5_1`) (b) Energy score over time (`FaultyCase5_Set5_1`)

Figure 5: Qualitative analysis of a successful detection: `FaultyCase5_Set5_1` (Slugging). (a) The energy distributions for normal (blue) and anomalous (red) data are well-separated, corresponding to a high ROC-AUC. (b) The energy score timeline, derived from all input features, shows a sharp, precise response. The red-shaded region indicates the ground-truth fault interval.



(a) Energy distribution (`FaultyCase2_Set2_1`) (b) Energy score over time (`FaultyCase2_Set2_1`)

Figure 6: Qualitative analysis of a challenging detection: `FaultyCase2_Set2_1` (Blockage). (a) The energy distributions are heavily overlapped, leading to a low ROC-AUC. (b) The energy score timeline, derived from all input features, shows a delayed and weak response. The red-shaded region indicates the ground-truth fault interval.

5 Discussion

5.1 Implications and Insights

Our results highlight several design principles for anomaly detection in complex industrial processes.

Representation matters as much as architecture. Performance on chaotic, high-frequency faults (e.g., FaultyCase 5) shows that transforming 1D signals into 2D time–frequency images via the CWT can convert a hard temporal problem into a textural pattern-recognition task, where Vision Transformers excel. Rather than replacing feature engineering, modern architectures benefit from representational choices that align with their inductive biases.

Architectural specialization is beneficial. Comparisons with the LSTM baseline suggest complementary strengths: recurrent models favor long-range, quasi-periodic patterns (FaultyCase 3), while our patch-based approach is strong for localized, non-stationary events (FaultyCase 5). This pattern argues for ensembles or hybrids that combine recurrent and attention-based components, rather than seeking a single model that dominates across all anomaly types.

EBMs separate anomaly scoring from decision-making. Cases with high ROC-AUC but lower F1 indicate that the energy function can be discriminative while a static decision threshold underperforms. This separation is useful: the EBM provides a principled anomaly score, and thresholding can be improved independently (e.g., via adaptive or sequential methods).

5.2 Limitations

Sensitivity to slow-drift anomalies. The most salient weakness is reduced sensitivity to slowly evolving faults (e.g., `FaultyCase2_Set2_1`, Section 4.4). Two factors contribute: (i) when detrending is applied, it can attenuate gradual deviations by design; and (ii) the ViT processes fixed-size chunks without recurrent memory, making subtle, long-horizon changes harder to detect. This manifests as delayed energy increases and overlapping score distributions (Figure 6).

Compute and hyperparameter sensitivity. CWT preprocessing and Energy Matching with a ViT are more resource-intensive than operating on raw signals, and performance depends on context length. As shown in our chunk-width analysis (Section 4.3), the optimal window balances modeling longer normal dynamics against sensitivity to short-lived faults, implying nontrivial tuning for new deployments.

Diagnostic interpretability. Energy maps localize anomalies in time–frequency space but do not attribute root-cause sensors directly. Because channels are concatenated and scored jointly, attributing a high-energy patch to specific inputs requires post-hoc analysis. Resolution is also bounded by the patch size, limiting sub-patch localization needed for fine-grained diagnosis.

5.3 Future Work

Hybrid models for multi-scale dynamics. Augment ScaloVit-EBM with a lightweight recurrent branch (e.g., LSTM/GRU) over raw or downsampled signals and fuse outputs, targeting robustness to both slow drifts and high-frequency events.

Attribution and hierarchical localization. Develop gradient-based or attention-driven attributions to relate high-energy regions to input channels, and trigger secondary, finer-grained analysis on high-energy patches to achieve sub-patch localization.

Adaptive and online deployment. Enable streaming operation with efficient, stateful CWT and adopt adaptive decision rules (e.g., change-point detection, Bayesian or percentile policies with drift correction) to maintain performance under nonstationarity.

6 Conclusion

This work presented ScaloVit-EBM, a localized energy-based model for unsupervised anomaly detection in multivariate industrial time series. By converting sensor streams into multichannel CWT scalograms and learning per-patch energies with a U-Net plus Vision Transformer backbone trained via Energy Matching on normal-only data, the method produces interpretable time–frequency energy maps and a principled anomaly score. A deployment-minded inference protocol—overlapping chunking with max aggregation and a percentile threshold calibrated on normal validation—keeps scoring and decision-making decoupled and simple to operate.

On a realistic three-phase flow benchmark, ScaloVit-EBM is competitive with or superior to strong reconstruction- and Transformer-based baselines, and excels on high-frequency slugging where localized time–frequency structure is critical. Ablations substantiate three practical design choices: detrending improves class separability; max aggregation avoids diluting localized deviations across overlaps; and chunk width trades off long-range context against sensitivity to short-lived events. Additional analyses show that localized, per-patch scoring is important for spatially complex faults, while augmenting the objective with Contrastive Divergence can improve F1 at the expense of ROC-AUC and training cost.

Limitations include reduced sensitivity to gradually evolving faults, dependence on chunk length and preprocessing, and limited channel-level attribution. These motivate extensions toward hybrid recurrent–attention architectures for multi-scale dynamics, hierarchical and channel-aware attribution mechanisms, and adaptive, streaming deployment to handle nonstationarity.

Acknowledgments

We would like to express our sincere gratitude to our main supervisor, Dr. Loo Junn Yong, and co-supervisor, Dr. Ting Chee Ming, for their invaluable guidance and support throughout this research. We also extend our thanks to Tew Hwa Hui for the fruitful discussions and cooperation. This work made use of the Three-Phase Flow Facility dataset, generously provided by Cranfield University [Ruiz-Cárcel et al., 2015]. The implementation of our energy-based model is adapted from the official Energy Matching codebase provided by Balcerak et al. in their public repository [Balcerak et al., 2025]. Furthermore, our signal preprocessing methodology was informed by the techniques presented in the ImagenTime project by Naiman et al. [2024].

References

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. Usad: Unsupervised anomaly detection on multivariate time series. page 3395–3404, Virtual Event, CA, USA, 2020. Association for Computing Machinery. doi: https://doi.org/10.1145/3394486.3403392. URL `https://doi.org/10.1145/3394486.3403392`.

Michal Balcerak, Tamaz Amiranashvili, Antonio Terpin, Suprosanna Shit, Lea Bogensperger, Sebastian Kaltenbach, Petros Koumoutsakos, and Bjoern Menze. Energy matching: Unifying flow matching and energy-based models for generative modeling, 2025. URL `https://arxiv.org/abs/2504.10612`.

Paul Boniol, Qinghua Liu, Mingyi Huang, Themis Palpanas, and John Paparrizos. Dive into time-series anomaly detection: A decade review. *arXiv (Cornell University)*, 12 2024. doi: 10.48550/arxiv.2412.20512.

Davide Carbone. Hitchhiker's guide on the relation of energy-based models with other generative models, sampling and statistical physics: a comprehensive review, 2025. URL `https://arxiv.org/abs/2406.13661`.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp, 2017. URL `https://arxiv.org/abs/1605.08803`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL `https://arxiv.org/abs/2010.11929`.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Neural Information Processing Systems*, 32:3603–3613, Jan 2019.

Pradeep Etikani, Vijaya Kamireddy, Krishnateja Shiva, Akhil Mittal, Arth Dave, Darshit Thakkar, Devidas Kanchetti, and Rajesh Munirathnam. Anomaly detection in sensor data with machine learning: Predictive maintenance for industrial systems, 01 2024.

Geoffrey E Hinton. *A practical guide to training restricted boltzmann machines*, page 599–619. Springer Berlin Heidelberg, 2012. doi: https://doi.org/10.1007/978-3-642-35289-8%E2%82%832. URL `https://doi.org/10.1007/978-3-642-35289-8%E2%82%832`.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions, 2018. URL `https://arxiv.org/abs/1807.03039`.

Yaron Lipman, Ricky T. Q., Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL `https://arxiv.org/abs/2210.02747`.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008. doi: 10.1109/ICDM.2008.17.

Qinghua Liu, Paul Boniol, Themis Palpanas, and John Paparrizos. Time-series anomaly detection: Overview and new trends. *Proceedings of the VLDB Endowment*, 17:4229–4232, 08 2024. doi: 10.14778/3685800.3685842. URL `https://inria.hal.science/hal-04708918/`.

Junn Yong Loo, Michelle Adeline, Julia Kaiwen Lau, Fang Yu Leong, Hwa Hui Tew, Arghya Pal, Vishnu Monn Baskaran, Chee-Ming Ting, and W Phan. Learning energy-based generative models via potential flow: A variational principle approach to probability density homotopy matching, 2025. URL `https://arxiv.org/abs/2504.16262`.

Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. Lstm-based encoder-decoder for multi-sensor anomaly detection, 2016. URL `https://arxiv.org/abs/1607.00148`.

Stephane G Mallat and Gabriel Peyre. *A wavelet tour of signal processing : the sparse way*. Academic Press, Cop, Amsterdam, 2009. ISBN 9780123743701.

N MARWAN, M CARMENROMANO, M THIEL, and J KURTHS. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5-6):237–329, Jan 2007. doi: https://doi.org/10.1016/j.physrep.2006.11.001.

I.P Mitov. A method for assessment and processing of biomedical signals containing trend and periodic components. *Medical Engineering & Physics*, 20(9):660–668, 1998. doi: https://doi.org/10.1016/S1350-4533(98)00077-0. URL `https://www.sciencedirect.com/science/article/pii/S1350453398000770`.

Mohsin Munir, Shoaib Ahmed Siddiqui, Andreas Dengel, and Sheraz Ahmed. Deepant: A deep learning approach for unsupervised anomaly detection in time series. *IEEE Access*, 7:1991–2005, 2019. doi: https://doi.org/10.1109/access.2018.2886457.

Ilan Naiman, Nimrod Berman, Itai Pemper, Idan Arbiv, Gal Fadlon, and Omri Azencot. Utilizing image transforms and diffusion models for generative modeling of short and long time series, 2024. URL `https://arxiv.org/abs/2410.19538`.

Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. *arXiv (Cornell University)*, Jan 2019. doi: https://doi.org/10.48550/arxiv.1904.09770.

A.V. Oppenheim and J.S. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5): 529–541, 1981. doi: https://doi.org/10.1109/proc.1981.12022.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2):1–38, Mar 2021. doi: https://doi.org/10.1145/3439950. URL `http://dx.doi.org/10.1145/3439950`.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. page 1530–1538, Lille, France, 2015. JMLR.org.

Manel Rhif, Ali Ben Abbes, Imed Farah, Beatriz Martínez, and Yanfang Sang. Wavelet transform application for/in non-stationary time-series analysis: A review. *Applied Sciences*, 9(7):1345, Mar 2019. doi: https://doi.org/10.3390/app9071345.

C. Ruiz-Cárcel, Y Cao, D Mba, L Lao, and R.T Samuel. Statistical process monitoring of a multiphase flow facility. *Control Engineering Practice*, 42:74–88, 2015. doi: https://doi.org/10.1016/j.conengprac.2015.04.012. URL `https://www.sciencedirect.com/science/article/pii/S0967066115000866`.

Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. volume 12, pages 582–588, 01 1999.

Chen Su, Zhengzhou Cai, Yuanhe Tian, Zhuochao Chang, Zihong Zheng, and Yan Song. Diffusion models for time series forecasting: A survey, 2025. URL `https://arxiv.org/abs/2507.14507`.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. 2019. ISBN 9781450362016. doi: https://doi.org/10.1145/3292500.3330672. URL `https://dl.acm.org/citation.cfm?id=3330672`.

David M.J. Tax and Robert P.W. Duin. Support vector data description. *Machine Learning*, 54:45–66, 01 2004. doi: 10.1023/b:mach.0000008084.60811.49.

Christopher Torrence and Gilbert P Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1):61–78, 1998. doi: https://doi.org/10.1175/1520-0477(1998)079%C2%A10061:APGTWA%C2%BF2.0.CO;2. URL https://journals.ametsoc.org/view/journals/bams/79/1/1520-0477%E2%82%81998%E2%82%8079%E2%82%80061%E2%82%90pgtwa%E2%82%82%E2%82%80%3Csub%3Ec%3C/sub%3Eo%E2%82%82.xml.

Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. Tranad. *Proceedings of the VLDB Endowment*, 15(6):1201–1214, Feb 2022. doi: https://doi.org/10.14778/3514061.3514067.

Samy-Melwan Vilhes, Gilles Gasso, and Mokhtar Z Alaya. Patchtrad: A patch-based transformer focusing on patch-wise reconstruction error for time series anomaly detection, 2025. URL https://arxiv.org/abs/2504.08827.

Zhiguang Wang and Tim Oates. Imaging time-series to improve classification and imputation, 2015. URL https://arxiv.org/abs/1506.00327.

Zhisheng Xiao, Qing Yan, and Yali Amit. Ebms trained with maximum likelihood are generator models trained with a self-adverserial loss, 2021. URL https://arxiv.org/abs/2102.11757.

Sangwoong Yoon, Young-Uk Jin, Yung-Kyun Noh, and Frank C Park. Energy-based models for anomaly detection: A manifold diffusion recovery approach. *arXiv (Cornell University)*, Jan 2023. doi: https://doi.org/10.48550/arxiv.2310.18677.

Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection, 2016. URL https://arxiv.org/abs/1605.07717.

Qianru Zhang, Peng Yang, Honggang Wen, Xinzhu Li, Haixin Wang, Fang Sun, Zezheng Song, Zhichen Lai, Rui Ma, Ruihua Han, Tailin Wu, Siu-Ming Yiu, Yizhou Sun, and Hongzhi Yi. Beyond the time domain: Recent advances on frequency transforms in time series analysis. Feb 2025. doi: https://doi.org/10.48550/arXiv.2504.07099.

Huimin Zhao, Shaoyan Zuo, Ming Hou, Wei Liu, Ling Yu, Xinhua Yang, and Wu Deng. A novel adaptive signal processing method based on enhanced empirical wavelet transform technology. *Sensors*, 18(10), 2018. doi: https://doi.org/10.3390/s18103323. URL https://www.mdpi.com/1424-8220/18/10/3323.

Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. 2018. URL https://openreview.net/forum?id=BJJLHbb0-.

Appendix