Enhancing Anomaly Detection in Oil and Gas Rotating Machinery through a Physics-Informed Data Filtration Pipeline

Gokul B

Abstract—**The reliability of anomaly detection models in the oil and gas sector is frequently compromised by the poor quality of raw sensor data from rotating machinery. This paper introduces a multi-stage, physics-informed data filtration pipeline designed to systematically remove operational noise and confounding artifacts from vibration data prior to model training. By applying this pipeline, the accuracy of an Isolation Forest-based anomaly detection model was significantly improved from a baseline of 56% on unfiltered data to a robust 92%. This enhancement transforms the system from an unreliable alerting mechanism into a trustworthy decision-support tool for predictive maintenance (PdM). The proposed framework represents a practical shift towards proactive maintenance strategies, which are critical for reducing unplanned downtime, minimizing operational costs, and enhancing safety in high-stakes industrial environments.**

Index Terms—**Anomaly Detection, Predictive Maintenance, Machine Learning, Vibration Analysis, Oil and Gas Industry, Data Filtration.**

I. Introduction

The oil and gas industry operates within a high-stakes environment where equipment reliability is critical. Failures in rotating machinery—such as pumps, compressors, and turbines—can lead to catastrophic financial, safety, and environmental consequences. These risks have accelerated the industry’s transition from traditional reactive and preventive maintenance strategies toward a data-driven predictive maintenance (PdM) paradigm [1].

PdM integrates real-time sensor data with machine learning (ML) algorithms to forecast faults before they occur. At its core lies *anomaly detection*—the process of identifying deviations from normal operational behavior that may signal impending failures [2]. However, the success of any ML model depends heavily on the quality of its input data. Harsh industrial environments often introduce noise, drift, and operational variability into sensor data, leading to unreliable models and frequent false alarms [3].

This paper directly addresses these data integrity challenges. We propose a physics-informed, multi-stage data filtration pipeline that systematically enhances vibration data before feeding it into an Isolation Forest (IF) anomaly detection model. Our results demonstrate a performance improvement from 56% to 92% accuracy, showing that informed preprocessing can be more impactful than algorithmic complexity alone.

II. Background and Related Work

Industrial maintenance strategies have evolved from reactive (repair after failure) to preventive (time-based maintenance), and now to predictive maintenance (condition-based) [**?**]. Predictive approaches leverage vibration signals as key indicators of machinery health. Every mechanical asset exhibits a characteristic vibration signature, and deviations often signify mechanical degradation such as bearing wear, shaft misalignment, or imbalance [4].

Machine learning enables automated analysis of vibration data, with unsupervised algorithms like One-Class SVM, Local Outlier Factor (LOF), and Isolation Forest (IF) performing well where labeled data are scarce [5]. However, most research focuses on algorithmic novelty rather than data preparation, leaving a gap between academic models and deployable industrial solutions.

Our work bridges this divide by demonstrating that physics-informed preprocessing—rooted in vibration analysis—can drastically improve anomaly detection robustness under variable operating conditions.

III. Proposed Framework

The proposed system comprises three major components: (1) raw data acquisition, (2) multi-stage physics-informed filtration, and (3) anomaly detection using Isolation Forest. Figure **??** illustrates the overall architecture.

A. Data Acquisition

Data were collected from centrifugal pumps and gas compressors operating in a refinery environment. Vibration data from accelerometers mounted on bearing housings were synchronized with tachometer-based rotational speed measurements. Raw data exhibited high-frequency noise and large amplitude fluctuations due to load variations, making direct ML application infeasible.

B. Stage 1: Spectral Transformation and Noise Reduction

The raw time-domain signal $s(n)$ is converted to its frequency-domain representation using the Fast Fourier Transform (FFT):

$$S(\omega) = \mathrm{FFT}\{s(n)\}. \tag{1}$$

This transformation allows clear separation between harmonic components (linked to mechanical faults) and broadband random noise. A high-pass filter removes low-frequency structural vibrations and sensor drift, improving the signal-to-noise ratio.

C. Stage 2: Rotational Speed Normalization (Order Tracking)

Rotating machines rarely operate at a constant speed. As speed changes, frequency components shift, causing spectral smearing. To normalize this, we define vibration features in terms of *orders* rather than absolute frequency:

$$\text{Order} = \frac{f}{f_r}, \tag{2}$$

where f_r is the instantaneous rotational frequency. This order-tracking technique ensures consistent fault signatures across variable speeds, aligning features that correspond to the same mechanical behavior.

D. Stage 3: Statistical Outlier Trimming and Feature Engineering

Transient spikes and sensor errors are filtered using the interquartile range (IQR) rule:

$$\text{Outlier if } x < Q_1 - 1.5 \cdot IQR \text{ or } x > Q_3 + 1.5 \cdot IQR, \tag{3}$$

where $IQR = Q_3 - Q_1$. This stage ensures model stability by removing non-repetitive disturbances. Finally, spectral-energy features are extracted for specific orders associated with typical fault modes (imbalance, bearing degradation, gear mesh, etc.).

E. Anomaly Detection using Isolation Forest

The Isolation Forest isolates anomalies by recursively partitioning the data. Points requiring fewer splits to isolate are likely anomalous. Given its efficiency and ability to handle high-dimensional datasets, IF is well-suited for real-time industrial environments [5].

IV. Experimental Validation

A. Dataset and Setup

The dataset comprises over 2,000 hours of labeled vibration and speed data, including known normal and fault conditions identified from maintenance logs. Approximately 95% of data corresponded to normal operation and 5% to anomalies.

Performance metrics include Accuracy, Precision, Recall, and F1-Score, defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \tag{5}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{6}$$

B. Results

Table I summarizes the incremental performance improvement as filtration stages are added.

V. Discussion

The progressive gains validate each stage's contribution: Stage 1 removes broadband noise, Stage 2 yields speed-invariant features, and Stage 3 provides statistical robustness. The final performance demonstrates that domain-aware preprocessing enables efficient models like IF to operate with industrial-grade reliability.

TABLE I: Performance Improvement with Data Filtration Stages

Model Config.	Accuracy	Precision	Recall	F1
Baseline (No Filter)	56.3%	58.1%	65.4%	0.615
+ Stage 1 (FFT+Filter)	71.8%	70.2%	75.1%	0.726
+ Stage 2 (Order Track)	88.5%	90.3%	87.9%	0.891
+ Stage 3 (Outlier Trim)	**92.1%**	**93.5%**	**91.2%**	**0.923**

VI. Conclusion and Future Work

We proposed a physics-informed multi-stage filtration pipeline for vibration-based anomaly detection. By transforming raw signals into speed-normalized and statistically robust features, the Isolation Forest model's accuracy improved from 56% to 92%. Future work includes adapting the pipeline to reciprocating machinery, implementing edge-compute real-time filtering, and integrating anomaly scores into a digital-twin for comprehensive asset health management.

Acknowledgment

The authors thank Bluespace AI for data access and computational resources.

References

[1] A. K. Singh and A. K. Sharma, "Predictive maintenance in the oil and gas industry: A review of ai-driven anomaly detection," *Journal of Petroleum Technology*, vol. 74, no. 3, pp. 88–95, 2022.

[2] M. Goldstein and S. Uchida, "A comparative evaluation of anomaly detection techniques for industrial systems," in *Proc. 2016 IEEE Symp. on Computational Intelligence and Data Mining (CIDM)*, 2016, pp. 1–8.

[3] D. Garcia and M. A. Rodriguez, "Data quality challenges in predictive maintenance for legacy industrial equipment," *Industrial Informatics Journal*, 2023, submitted.

[4] H. P. Bloch and F. K. Geitner, *Machinery Failure Analysis and Troubleshooting*, 4th ed. Gulf Professional Publishing, 2012.

[5] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 2008 Eighth IEEE Int. Conf. on Data Mining*, 2008, pp. 413–422.