Learning Rewards, Not Labels: Adversarial Inverse Reinforcement Learning for Machinery Fault Detection

Extended Abstract

Dhiraj Neupane
Deakin University
Waurn Ponds, Victoria, Australia
d.neupane@deakin.edu.au

Richard Dazeley
Deakin University
Waurn Ponds, Victoria, Australia
richard.dazeley@deakin.edu.au

Mohamed Reda Bouadjenek
Deakin University
Waurn Ponds, Victoria, Australia
reda.bouadjenek@deakin.edu.au

Sunil Aryal
Deakin University
Waurn Ponds, Victoria, Australia
sunil.aryal@deakin.edu.au

ABSTRACT

Reinforcement learning (RL) offers significant promise for machinery fault detection (MFD). However, most existing RL-based MFD approaches do not fully exploit RL's sequential decision-making strengths, often treating MFD as a simple *guessing game* (Contextual Bandits). To bridge this gap, we formulate MFD as an offline inverse reinforcement learning problem, where the agent learns the reward dynamics directly from healthy operational sequences, thereby bypassing the need for manual reward engineering and fault labels. Our framework employs *Adversarial Inverse Reinforcement Learning* to train a discriminator that distinguishes between normal (expert) and policy-generated transitions. The discriminator's learned reward serves as an anomaly score, indicating deviations from normal operating behaviour. When evaluated on three run-to-failure benchmark datasets (HUMS2023, IMS, and XJTU-SY), the model consistently assigns low anomaly scores to normal samples and high scores to faulty ones, enabling early and robust fault detection. By aligning RL's sequential reasoning with MFD's temporal structure, this work opens a path toward RL-based diagnostics in data-driven industrial settings.

KEYWORDS

Adversarial Learning; Anomaly Detection; Machinery Fault Detection; Reinforcement Learning; Prediction

ACM Reference Format:
Dhiraj Neupane, Richard Dazeley, Mohamed Reda Bouadjenek, and Sunil Aryal. 2026. Learning Rewards, Not Labels: Adversarial Inverse Reinforcement Learning for Machinery Fault Detection: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/AXYX4522

1 INTRODUCTION

Machinery fault detection (MFD) is essential for maintaining industrial reliability, yet acquiring extensive labelled fault data remains a major bottleneck. While supervised learning dominates the field (≈81% of studies [6]), it suffers significantly from the scarcity of fault labels in real-world settings. Reinforcement learning (RL) has emerged as a promising alternative, aiming to model the sequential nature of degradation. However, most existing RL-based MFD approaches [1, 3, 11] reduce the problem to a static *guessing game* or *contextual-bandit* (CB) task. In these setups, agents treat sensor samples as independent states, issue one-shot classification actions, and ignore the discount factor ($\gamma = 0$), thereby discarding the temporal structure inherent in fault progression [7].

This simplification violates RL's core premise of sequential decision-making. To bridge this gap, we propose formulating MFD as an offline *Inverse Reinforcement Learning* (IRL) problem. Unlike standard RL, which requires a manually specified reward function, a significant challenge in complex machinery, IRL learns the reward function directly from expert demonstrations [8].

We introduce an *Adversarial Inverse Reinforcement Learning* (AIRL) [2] framework that learns the reward dynamics of *healthy* machine operation. By treating normal operational sequences as "expert" trajectories, our discriminator learns to distinguish between healthy transitions and generated anomalies. The learned reward function acts as an interpretable anomaly score: high rewards indicate alignment with healthy dynamics, while low rewards signal deviations. To the best of our knowledge, this is the first application of AIRL to MFD. Extensive experiments on three run-to-failure benchmarks (HUMS2023 [17], IMS [12], XJTU-SY [16]) demonstrate that our framework enables early and robust fault detection without requiring fault labels, outperforming traditional one-class and reconstruction-based baselines.

2 METHODOLOGY

We formulate MFD as an *offline IRL problem* where the goal is to recover a reward function that rationalizes the behavior of a healthy (normal) machinery (the *expert*).

2.1 State Transition Construction

Since industrial fault datasets lack recorded control inputs, we adopt a *State-Only Imitation Learning (SOIL)* formulation [14]. We segment the normalized vibration signals into fixed-length windows. Because explicit control actions are absent, to apply Inverse RL in

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/AXYX4522

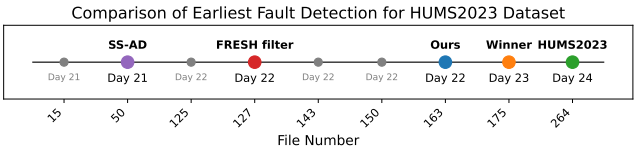

Figure 1: Earliest Detection Onset Comparison (HUMS2023)

this action-free setting, we define the state s_t as the current window and treat the *system's natural temporal evolution* to the next window as a "proxy action" ($a_t = x_{t+1}$). This formulation allows the AIRL discriminator to evaluate the plausibility of the transition dynamics ($s_t \to s_{t+1}$) by scoring them against the healthy expert distribution.

2.2 Adversarial Reward Learning

We employ AIRL [2], which frames reward learning as a GAN-like optimization with two components: a *Generator (π)* trained to mimic expert dynamics, and a *Discriminator (D)*.

The discriminator $D(s, a, s')$ estimates the probability that a transition is from the healthy expert distribution rather than the generator. Crucially, to recover a meaningful signal, AIRL structures the discriminator as:

$$D(s, a, s') = \sigma\left(r_\theta(s, a) + \gamma V_\phi(s') - V_\phi(s) - \log \pi(a|s)\right) \quad (1)$$

where σ is the sigmoid function. This structural constraint forces the learned term $r_\theta(s, a)$ to act as a robust reward function (or health score), disentangled from the system dynamics.

2.3 Anomaly Scoring

Once trained, the discriminator D estimates the probability that a transition belongs to the healthy manifold. High values indicate alignment with expert dynamics, while low values signal "surprising" deviations. We quantify this by defining the anomaly score for a trajectory τ as the inverted average discriminator confidence:

$$\text{Score}(\tau) = 1 - \frac{1}{T}\sum_{t=0}^{T} D(s_t, a_t, s_{t+1}) \quad (2)$$

Fault onset is then identified by thresholding this score using dynamic methods (e.g., Otsu's method [9], K-means [5]) and standard statistical rules.

3 EXPERIMENTS

3.1 Experimental Setup

We evaluated the framework on three run-to-failure benchmark datasets: *HUMS2023* (helicopter gearbox fatigue), *IMS*, and *XJTU-SY*. For the primary HUMS2023 dataset, we selected the *Ring-Front 2 (RF2)* accelerometer, as it best captures the fault-sensitive gear-meshing dynamics. The models were trained exclusively on healthy data (Days 17–20) and tested on the degradation phase (Days 21–27). We compared AIRL against standard baselines (Isolation Forest (IF) [4], one-class support vector machines (OCSVM) [13]), reconstruction models (Autoencoder (AE), Variational Autoencoder (VAE)), and temporal reconstruction methods (LSTM-AE, LSTM-VAE). Additionally, we benchmarked against recent state-of-the-art methods including *SS-AD* [6] and the *FRESH-filter* [15], as well as a *Contextual Bandit (CTQN)* [3] baseline representing current RL-based MFD approaches.

Table 1: Earliest fault detection on HUMS2023.

Model	Detection	Model	Detection
IF	Day 21 (#9)	VAE	Day 21 (#37)
OCSVM	Day 21 (#37)	LSTM-VAE	Day 22 (#131)
AE	Day 21 (#1)	SS-AD [6]	Day 21 (#50)
LSTM-AE	Day 22 (#131)	FRESH filter [15]	Day 22 (#127)
CTQN (CB)	*No Fault*	*CW* [10]	*Day 23 (#175)*
Committee GT	*Day 24 (#264)*	**AIRL (Ours)**	**Day 22 (#163)**

GT: Ground Truth; CW: Challenge Winner

3.2 Results

The primary evaluation metric was the *earliest valid detection* of fault onset. Due to space constraints, we detail the detection results for the primary HUMS2023 dataset in Table 1; however, consistent performance trends were observed across the IMS and XJTU-SY benchmarks.

Our *AIRL framework* identified the fault onset at *Day 22 (File #163)*. As visualized in Figure 1, this detection falls squarely between the FRESH filter (File #127) and the official *Challenge Winner* (Day 23, File #175) [10]. Crucially, our detection precedes the conservative ground truth established by the HUMS committee (Day 24, File #264), demonstrating that AIRL provides a valuable early warning window without the premature false positives observed in other methods. Beyond timely detection, AIRL demonstrated superior *post-Detection Consistency (PDC)*, maintaining a steady anomaly rate (≈65%) after fault onset. This stability was mirrored in the IMS and XJTU-SY experiments, confirming the robustness of the sequential reward formulation.

In comparison, as shown in Table 1, standard baselines (IF, OCSVM, AE) flagged anomalies prematurely. While sequential models like LSTM-AE and LSTM-VAE improved precision (Day 22), they still triggered earlier than our method. Crucially, the Contextual Bandit (CTQN) baseline *failed entirely*, classifying the whole test set as normal. This confirms that without modeling state transitions ($\gamma = 0$), the agent cannot perceive the gradual accumulation of fatigue damage.

4 CONCLUSION

This work introduces the first application of Adversarial Inverse Reinforcement Learning to machinery fault detection. Unlike existing "RL-based" methods that treat fault diagnosis as a static *contextual-bandit* problem, our framework respects the sequential nature of machine degradation. By recovering a reward function from healthy data, AIRL provides a robust, interpretable anomaly score that detects faults early and reliably. Our results on HUMS2023, IMS, and XJTU-SY demonstrate that learning the *dynamics* of health is superior to merely classifying isolated observations. Future work will extend this framework to multi-sensor fusion and incorporate uncertainty-aware thresholding to further reduce false alarms in variable operating conditions.

REFERENCES

[1] Yu Ding, Liang Ma, Jian Ma, Mingliang Suo, Laifa Tao, Yujie Cheng, and Chen Lu. 2019. Intelligent fault diagnosis for rotating machinery using deep Q-network based health state classification: A deep reinforcement learning approach. *Advanced Engineering Informatics* 42 (2019), 100977.

[2] Justin Fu, Katie Luo, and Sergey Levine. 2018. Learning Robust Rewards with Adversarial Inverse Reinforcement Learning. In *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=rkHywl-A-

[3] Zhenning Li, Hongkai Jiang, and Yutong Dong. 2025. A convolutional-transformer reinforcement learning agent for rotating machinery fault diagnosis. *Expert Systems with Applications* 271 (2025), 126669.

[4] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 413–422.

[5] James B McQueen. 1967. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.* 281–297.

[6] Dhiraj Neupane, Mohamed Reda Bouadjenek, Richard Dazeley, and Sunil Aryal. 2024. A Comparative Study of Semi-Supervised Anomaly Detection Methods for Machine Fault Detection. In *PHM Society European Conference*, Vol. 8. 10–10.

[7] Dhiraj Neupane, Mohamed Reda Bouadjenek, Richard Dazeley, and Sunil Aryal. 2024. Machinery Fault Detection using Advanced Machine Learning Techniques. In *PHM Society European Conference*, Vol. 8. 4–4.

[8] Andrew Y Ng, Stuart Russell, et al. 2000. Algorithms for inverse reinforcement learning.. In *Icml*, Vol. 1. 2.

[9] Nobuyuki Otsu et al. 1975. A threshold selection method from gray-level histograms. *Automatica* 11, 285-296 (1975), 23–27.

[10] Cédric Peeters, Wenyi Wang, David Blunt, Timothy Verstraeten, and Jan Helsen. 2024. Fatigue crack detection in planetary gears: Insights from the HUMS2023 data challenge. *Mechanical Systems and Signal Processing* 212 (2024), 111292.

[11] Gensheng Qian and Jingquan Liu. 2022. Development of deep reinforcement learning-based fault diagnosis method for rotating machinery in nuclear power plants. *Progress in Nuclear Energy* 152 (2022), 104401.

[12] Hai Qiu, Jay Lee, Jing Lin, and Gang Yu. 2006. Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics. *Journal of sound and vibration* 289, 4-5 (2006), 1066–1090.

[13] Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. 1999. Support vector method for novelty detection. *Advances in neural information processing systems* 12 (1999).

[14] Faraz Torabi, Garrett Warnell, and Peter Stone. 2019. Adversarial imitation learning from state-only demonstrations. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 2229–2231.

[15] Rik Vaerenberg, Alex Ricardo Mauricio, and Konstantinos Gryllias. 2025. Detecting planet gear crack propagation using FRESH filters. In *14th Defence Science & Technology (DST) International Conference on Health and Usage Monitoring HUMS2025 Proceedings*.

[16] Biao Wang, Yaguo Lei, Naipeng Li, and Ningbo Li. 2020. A Hybrid Prognostics Approach for Estimating Remaining Useful Life of Rolling Element Bearings. *IEEE Transactions on Reliability* 69, 1 (2020), 401–412. https://doi.org/10.1109/TR.2018.2882682

[17] Wenyi Wang, David Blunt, and J Kappas. 2023. Helicopter main gearbox planet gear crack propagation test dataset.