Explainable AI for Maintenance Decision Support of Cryogenic Hydrogen Pumps

Rocco Cassandro[a,b], Katrina M. Groth[b,c]
[a] Systems Risk and Reliability Analysis Lab (SyRRA), Center for Risk and Reliability,
University of Maryland, College Park, MD, 20742 USA
[b] rcas@umd.edu
[c] kgroth@umd.edu

Abstract: Hydrogen fueling stations are critical infrastructures for enabling large-scale hydrogen mobility. As station capabilities increase, particularly with cryogenic liquid hydrogen (LH_2) storage, reliable operation of LH_2 pumps becomes essential for safety, reliability, and overall station efficiency. However, the high-dimensional, time-dependent, and nonlinear characteristics of pump sensor data, combined with the need for interpretable and actionable predictions, limit the effectiveness of traditional machine learning approaches for fault diagnostics and maintenance planning. Building on these challenges, this paper explores an explainable AI-based decision support framework for condition-based maintenance of LH_2 pumps. The framework processes multivariate time-series data using physics-informed feature extraction and degradation-sensitive indicators, feeding predictive classification models based on ensemble learning. SHapley Additive exPlanations (SHAP) is integrated for global and local feature attribution, linking explainable predictions directly to maintenance decisions. Feature attribution analysis reveals physically consistent degradation signatures aligned with known failure modes, supporting transparent, evidence-based decision-making under uncertainty. Overall, the framework advances trustworthy AI for safety-critical cryogenic systems and provides a scalable methodology for actionable predictive maintenance decisions in hydrogen infrastructure.

1. INTRODUCTION

Liquid hydrogen (LH_2) reciprocating pumps play a central role in hydrogen fueling infrastructure, particularly in stations employing LH_2 storage to support high dispensing capacities and efficient compression processes [1, 2]. During operation, these pumps are subjected to demanding conditions, including cryogenic temperatures, elevated pressures, and repeated mechanical loading, which contribute to diverse and interacting degradation phenomena [3]. As a result, maintaining reliable pump performance remains a significant challenge, with reported failure rates exceeding those observed in comparable oil and gas pumping systems. Pump failures can reduce station availability, interrupt hydrogen delivery, and increase the potential for hazardous hydrogen leaks [2, 4]. These challenges highlight the need for robust condition monitoring approaches capable of supporting safe and dependable operation of LH_2 fueling systems.

To address these operational and safety concerns, prognostics and health management (PHM) frameworks have gained increasing attention as tools for monitoring system health and enabling condition-informed maintenance strategies. In the literature, LH_2 pump reliability has primarily been assessed using physics-based methods, including finite element and computational fluid dynamics simulations, failure mode and effect analysis, and physics-of-failure models that account for fatigue, cavitation, and cryogenic material behavior [5, 6]. While these approaches provide valuable mechanistic insights, they often rely on assumptions and struggle to capture the variability of real-world operations.

Data-driven methods leverage sensor measurements to identify degradation trends directly from operational data. Within machine learning and artificial intelligence (ML and AI), deep learning techniques have demonstrated strong capabilities in pattern recognition [7]. However, their

effectiveness typically depends on large, well-labeled datasets that are rarely available in real-world systems due to sparse failure events, high testing costs, and limited operational histories, reflecting a typical challenge encountered in reliability engineering real-world problems [8]. Unsupervised learning approaches, including clustering and change-point detection, alleviate the dependence on labeled data by identifying patterns [9]. While these methods can reveal degradation behavior, the resulting outputs often lack direct diagnostic interpretation and therefore provide limited support for actionable maintenance decisions. In recent years, literature has further emphasized the importance of integrating PHM outputs with probabilistic risk assessment frameworks for complex engineering systems [10]. Such frameworks rely on accurate diagnostics and health state estimation as foundational elements for supporting risk-informed decision-making. Consequently, developing interpretable health state diagnostic models is a crucial step toward more comprehensive PHM capabilities for hydrogen fueling infrastructure, ultimately enabling efficient maintenance strategies.

This paper presents an xAI-based framework for health state diagnosis of LH_2 pumps operating in a real hydrogen fueling station. The proposed approach extracts run-level features from multivariate sensor data, including statistical descriptors and temporal indicators of system behavior. Ensemble learning models are trained using health state labels derived from degradation-stage segmentation of the learned health index trajectories. To improve explainability, SHAP (SHapley Additive exPlanations) is employed to quantify feature contributions at both global and local levels, supporting actionable diagnostic insights. The main contributions of this work are: (i) the integration of explainable machine learning techniques to improve transparency and trust in PHM model predictions, (ii) the development of a data-driven health state classification framework tailored to LH_2 pumps under limited and noisy data conditions, and (iii) the demonstration of the approach on real operational data from a hydrogen fueling station, highlighting its applicability for condition-based maintenance.

2. SYSTEM OVERVIEW AND OPERATIONAL VARIABLES

The system under study is a LH_2 fueling station in which a cryogenic reciprocating pump acts as the primary compression component. LH_2 is stored in a vacuum-insulated tank, where heat ingress generates boil-off gas (BOG) that must be vented or recovered [11]. The pump compresses LH_2 prior to vaporization into gaseous hydrogen (GH_2), which is subsequently stored in low- and high-pressure tanks for dispensing. To improve operational efficiency, vapor generated within the pump is recirculated to the storage tank, while a parallel GH_2 compressor recompresses part of the BOG for reintegration into the system.

This study focuses on three major failure modes affecting pump operation [9]. The first, Fail to Operate (FTO), occurs when the pump does not start or respond when it is required to. The second, Abnormal High Vapor (AHV), involves excessive hydrogen vapor generation in the cold end, which increases vapor return flow to the cryogenic storage tank. The third, Abnormal Low Output (ALO), is characterized by a reduced discharge flow rate and an inability of the system to reach its intended maximum pressure levels. Detection of these failure modes is based on the observation of multiple measurements collected from different parts of the system.

The fueling station is instrumented with a diverse set of sensors spanning pump operation, LH_2 storage, GH_2 compression, gaseous storage, vaporization, and overall system status. Measurements include electrical signals, temperatures, pressures, valve states, flow indicators, and operational metadata. In total, 95 raw variables are acquired at regular time intervals and subsequently processed into run-level feature representations capturing statistical, temporal, and frequency-domain characteristics of system behavior. A summary of the operational variables is provided in Table 1.

Table 1: Operational variables of hydrogen fueling station under study

Subsystem / Group	Variable	Description	Unit	Type
Pump Electrical	Current	Current drawn by pump motor, indicative of mechanical load	A	Continuous
	Frequency	Operating frequency of the pump drive	Hz	Continuous
Pump Temperatures	Temperature_1–4	Temperature sensors at seals, bearings, cold end	F	Continuous
Pump Pressures	Pressure	Pressure within pump compression chamber	psi	Continuous
Pump Valves	Valve_1–3	Operational state of pump valves controlling flow	Binary (0/1)	Discrete
Pump Diagnostics	Cavitation_Index	Indicator of cavitation severity based on pressure/flow instability	Binary (0/1)	Discrete
Liquid Tank Subsystem	Tank_Level	LH_2 fill level in storage tank	in-H2O	Continuous
	Tank_Pressure	Internal tank pressure	psi	Continuous
	Tank_Valve_1–2	Valve states controlling inflow/outflow	Binary (0/1)	Discrete
Compressor Subsystem	Crankcase_Temperature	Compressor crankcase temperature	F	Continuous
	Discharge_Pressure	Pressure at compressor outlet	psi	Continuous
	Discharge_Temperature	Gas temperature at outlet	F	Continuous
	Interstage_Pressure	Pressure between compression stages	psi	Continuous
Gaseous Storage Subsystem	Pressure_1–2	LP/HP GH_2 tank pressures	psi	Continuous
	Valve	Valve controlling gas flow	Binary (0/1)	Discrete
Vaporizers	Pressure	Pressure within vaporizer during phase change	psi	Continuous
	Temperature_1–2	Temperature at vaporizer stages	F	Continuous

3. METHODOLOGY

This work proposes an end-to-end explainable condition-based maintenance (CBM) framework for LH_2 pumps operating in hydrogen fueling stations. The methodology integrates multivariate operational data acquisition, run-level structuring, degradation-state labeling, subsystem-informed feature engineering, ensemble tree-based classification, and SHAP-based explainability into a unified workflow. The overall objective is to preserve physical interpretability while enabling identification of pump health states and failure modes under non-stationary operating conditions.

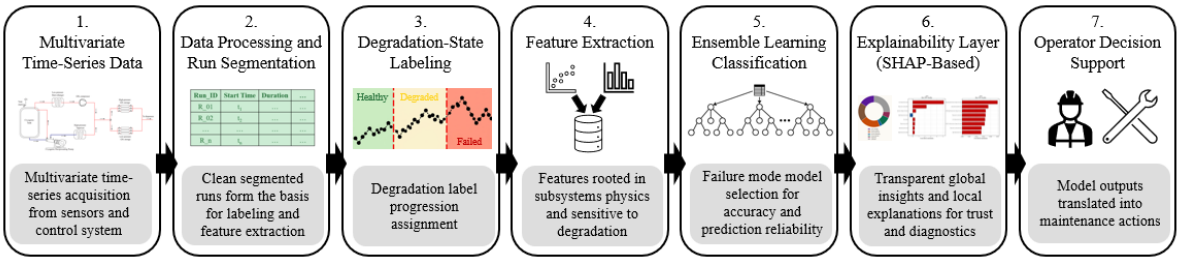

Figure 1: Proposed methodology

3.1. Multivariate Time-Series Data Acquisition

At each time instant t, the system is represented by a multivariate observation vector $x(t) = [x_1(t), x_2(t), \dots, x_m(t)]^{\mathrm{T}}$ where each component corresponds to a measured variable such as pump pressure, pump temperature, motor current, LH_2 tank pressure and level, vaporizer temperature, GH_2 storage pressure, cavitation state, or control-related variables. The resulting multivariate time series is non-stationary due to intermittent pump usage, varying duty cycles, and changing operating conditions

across refueling events. Consequently, the data are represented as a sequence of operational trajectories associated with individual pump runs, reflecting the non-stationary nature of the operating conditions.

3.2. Run-Level Structuring of Operational Data

To preserve operational meaning, the continuous monitoring stream is transformed into a run-based dataset. Let $\mathcal{D} = \{X^{(1)}, X^{(2)}, \dots, X^{(N)}\}$ denote the collection of run-level samples, where each $X^{(i)} \in \mathbb{R}^{T_i \times m}$ represents the multivariate evolution of the system during the i-th pump run. In the implemented workflow, each run is indexed by a pump-run identifier and filtered to retain only meaningful cycles, while short or low-current transients are excluded to reduce contamination from startup and shutdown behavior. This run-level formulation converts raw time-series data into a format suitable for degradation modeling and supervised learning.

3.3. Degradation-State Formulation

A central challenge in LH_2 pump degradation modeling is that maintenance records provide only the failure mode, while the underlying evolution of pump health remains unobserved. In addition, pump degradation occurs gradually and under highly variable operating conditions, making it difficult to define a direct health metric. To address this, latent health states are inferred from the learned health index (HI) trajectory and segmented into three ordered degradation stages representing healthy, incipient degradation, and degraded operation. Bayesian offline change-point segmentation with dynamic programming is used to identify the optimal breakpoints under a Normal-Inverse-Gamma probabilistic model while enforcing minimum segment-length constraints to ensure stable and non-degenerate state assignments. This approach can be found in [9]. The output consists of ordinal health labels, which serve as targets for downstream supervised learning.

3.4. Subsystem-Informed and Degradation-Sensitive Feature Representation

Each run-level sample $X^{(i)}$ is mapped to a feature vector $f^{(i)} = \phi\left(X^{(i)}\right) \in \mathbb{R}^q$, where $\phi(\cdot)$ aggregates multivariate sensor trajectories into engineered run-level descriptors. The feature representation combines statistical summaries and subsystem-informed variables derived from process variables. This results in a representation that includes signal-level summaries (e.g., range, linear slope, net change, percentiles, and, spike count, autocorrelation, and zero-crossing rate), health state-based means and slopes with deltas, derived physical combinations such as storage pressure span, vaporizer temperature span, pump pressure per amp, and pump pressure per hertz, and threshold-driven event metrics for cavitation, LH_2 venting, pump speed changes, pump pressure, and pressure reversals. The feature set also incorporates robust baseline-deviation descriptors computed as z-scores and median ratios. The features associated with direct control signals are excluded, ensuring the model emphasizes process behavior rather than operator commands

3.5. Ensemble Learning for Health State Classification

Each sample corresponds to a single pump operating run. The run-level feature vector $f^{(i)} \in \mathbb{R}^q$ summarizes the behavior of the system during that run through engineered descriptors described in Section 3.4. The objective of the ensemble learning is to map these operational signatures into discrete health states that characterize degradation progression for a specific failure mode scenario.

The extracted run-level feature vector $f^{(i)} \in \mathbb{R}^q$ is used to predict the ordinal health state $y^{(i)} \in \{0,1,2\}$. The classifier is defined as

$$\hat{y}^{(i)} = \underset{c \in \{0,1,2\}}{\operatorname{argmax}} \hat{P}(y^{(i)} = c \mid f^{(i)}), \tag{1}$$

where $\hat{P}(y = c \mid f)$ is the estimated posterior probability of class c. Tree-based ensemble learning is employed due to its ability to capture nonlinear decision boundaries, interactions among subsystem variables, and threshold-driven operational behavior.

In general, ensemble learning aims to combine multiple base learners $h_k(\cdot)$ into a composite predictor [12]. A generic additive form can be expressed as

$$F(f) = \sum_{k=1}^{K} \alpha_k \, h_k(f), \tag{2}$$

where $h_k(f)$ denotes the output of the k-th base learner and α_k its weight. This formulation is most general and is particularly representative of boosting-based ensembles. Bagging-based methods correspond to the special case of uniform weighting and probability averaging [13].

For multiclass boosting models, a class-specific score can be defined as

$$F_c(f) = \sum_{k=1}^{K} \alpha_k \, h_{k,c}(f), \tag{3}$$

where $h_{k,c}(f)$ is the contribution of the k-th learner to class c. The corresponding posterior probability is obtained via the softmax function

$$\hat{P}(y = c \mid f) = \frac{\exp(F_c(f))}{\sum_{r=0}^{2} \exp(F_r(f))}. \tag{4}$$

The learning objective is to minimize the empirical multiclass loss over the run-level dataset

$$L(\Theta) = \sum_{i=1}^{N} l\left(y^{(i)}, \hat{p}^{(i)}\right), \tag{5}$$

where Θ denotes model parameters and $\hat{p}^{(i)}$ is the predicted class-probability vector for run i. For probabilistic multiclass ensembles, the loss is typically expressed as cross-entropy

$$l\left(y^{(i)}, \hat{p}^{(i)}\right) = -\sum_{c=0}^{2} I\left(y^{(i)} = c\right) \log \hat{P}(y^{(i)} = c \mid f^{(i)}). \tag{6}$$

The output of this stage is a probabilistic health state classifier that assigns each pump run to one of three degradation states. The resulting class probabilities constitute the primary diagnostic output used for the subsequent explainability analysis.

3.5.1 Bagging-Based Ensembles: Random Forest and Extra Trees

Bagging-based ensembles, including Random Forest (RF) and Extra Trees (ET), construct multiple decision trees independently and aggregate their predicted class probabilities to improve stability and reduce variance. The aggregated prediction is given by

$$\hat{P}(y = c \mid f) = \frac{1}{K} \sum_{k=1}^{K} \hat{P}_k(y = c \mid f). \tag{7}$$

Random Forest trains each tree on a bootstrapped sample of the data. At each node, a subset of features is randomly selected, but the optimal split is chosen to maximize impurity reduction. Extra Trees introduce additional randomness by selecting split points randomly rather than optimizing for the best split [14]. Trees may be trained on the whole dataset or bootstrapped samples. This approach increases tree diversity and speeds up training. While both RF and ET reduce variance effectively, they do not leverage sequential learning and therefore may be less sensitive to distributed degradation patterns.

The output of RF and ET is a run-level probability distribution over health states derived from an ensemble voting mechanism across multiple trees, resulting in predictions that are robust to changes in operating conditions.

3.5.2 Boosting-Based Ensembles: Histogram Gradient Boosting and XGBoost

Boosting-based ensembles, such as Histogram Gradient Boosting (HistGB) and XGBoost, construct trees sequentially, with each new tree aiming to correct the residual errors of the previous ensemble

$$J^{(t)} = \sum_{i=1}^{N} l\left(y^{(i)}, F^{(t-1)}\left(f^{(i)}\right) + h_t\left(f^{(i)}\right)\right) + \Omega(h_t), \tag{8}$$

where $F^{(t-1)}$ is the ensemble prediction from the previous iteration, h_t is the newly added tree, and $\Omega(h_t)$ is a regularization term controlling model complexity.

Histogram Gradient Boosting accelerates training by discretizing continuous features into histograms for efficient split finding. Its sequential structure makes it effective for capturing incipient degradation patterns distributed across multiple variables [15]. XGBoost extends gradient boosting by incorporating second-order derivative information for optimization and applying L1/L2 regularization, improving generalization and reducing overfitting [16]. It is particularly effective when degradation signatures are nonlinear and strongly interdependent. Boosting methods primarily reduce bias by sequentially correcting residual errors, making them well suited for gradual or incipient transitions between health states.

The output of the boosting models is a refined probabilistic health state prediction that integrates nonlinear relationships across multiple variables, aimed at improving the detection of failure mode scenarios degradation progressions.

3.5.3 Tree Splitting and Impurity Reduction

At each internal node, candidate splits are evaluated to reduce class impurity. Using the Gini criterion, the impurity of node n is

$$G(n) = 1 - \sum_{c=0}^{2} p(c \mid n)^2, \tag{9}$$

where $p(c \mid n)$ is the proportion of class $c$at node n. The optimal split maximizes impurity reduction

$$\Delta G = G(n) - \frac{N_L}{N_n} G(n_L) - \frac{N_R}{N_n} G(n_R), \tag{10}$$

where n_L and n_R are the left and right child nodes, and N_L, N_R, N_n are the corresponding sample counts. This process produces decision structures that partition the operational feature space into regions associated with different degradation states.

3.5.4 Model Evaluation

Given the class imbalance across health states, model assessment is performed with macro F1 score

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{3}\sum_{c=0}^{2} \mathrm{F1}_c, \tag{11}$$

to prevent overestimation of performance due to dominance of the healthy class. In the implemented workflow, multiple ensemble models are evaluated separately for each failure mode, and the final deployed classifier is selected on a failure mode-specific basis according to predictive performance.

This evaluation stage identifies the ensemble model that provides the most accurate classification and consistent performance across health states for a given failure mode.

3.6. Explainable Decision Support via SHAP

While accurate classification is important, practical CBM also requires understanding why a particular health state prediction was made. To translate model outputs into actionable engineering decisions, the proposed framework employs SHapley Additive exPlanations (SHAP) [17], which provide a locally additive decomposition of the class score for a given run-level feature vector f. For a target class c, the model output is represented as

$$F_c(f) = \phi_{0,c} + \sum_{j=1}^{q} \phi_{j,c}, \tag{12}$$

where $\phi_{0,c}$ is the baseline contribution for class c, and $\phi_{j,c}$ is the contribution of feature j to the class-specific prediction.

The SHAP value for feature j is defined from Shapley value theory as

$$\phi_{j,c} = \sum_{S \subseteq \mathcal{F} \setminus \{j\}} \frac{|S|!(q-|S|-1)!}{q!} [v_c(S \cup \{j\}) - v_c(S)], \tag{13}$$

where $\mathcal{F} = \{1, \ldots, q\}$ is the full feature set and $v_c(S)$ is the value function associated with subset S for class c. Intuitively, $\phi_{j,c}$ measures the marginal contribution of feature j averaged over all possible coalitions of the remaining features.

For tree ensembles, SHAP values can be computed efficiently using TreeSHAP, which exploits tree structure to avoid exhaustive coalition enumeration. This is especially suitable here because the predictive models are tree-based. Global feature importance is then obtained by averaging absolute SHAP magnitude across all runs

$$I_j = \frac{1}{N} \sum_{i=1}^{N} \left| \phi_{j,c^{(i)}}^{(i)} \right|. \tag{14}$$

This produces a ranking of features according to their average influence on model output. However, feature-level explanations alone may be difficult to interpret from an engineering standpoint, particularly when multiple derived indicators originate from the same physical subsystem. Therefore, features are grouped into subsystem-level sets G_k, and subsystem attribution is computed as

$$\Phi_k^{(i)} = \sum_{j \in G_k} \phi_{j,c^{(i)}}^{(i)}. \tag{15}$$

Similarly, a global subsystem importance can be defined as

$$\overline{\Phi}_k = \frac{1}{N} \sum_{i=1}^{N} \left| \Phi_k^{(i)} \right|. \tag{16}$$

This aggregation enables direct mapping from model reasoning to physical components such as pump hydraulics, tank dynamics, vaporizer behavior, GH_2 storage response, pump thermal loading, and operational controls. For a given run i, let the predicted class be

$$\hat{c}^{(i)} = \underset{c}{\operatorname{argmax}}\, \hat{P}(y = c \mid f^{(i)}). \tag{17}$$

The local explanation is then formed using the SHAP vector corresponding to $\hat{c}^{(i)}$, allowing the prediction to be interpreted through the dominant positive and negative contributors that drove the model toward that state. This provides a transparent bridge between learned statistical behavior and

operator-facing maintenance reasoning. This local-global explanation structure is especially important in CBM settings, where the objective is not only accurate classification, but also actionable and physically interpretable diagnostic support.

The output of this stage is a set of local and global diagnostic explanations that identify the operational variables and physical subsystems driving each health state prediction. These explanations support trustworthy condition-based maintenance decisions and provide actionable engineering insight for managing degradation within station operations.

4. RESULTS AND DISCUSSION

The models with direct control variables removed maintained strong multiclass health-state classification performance across all three failure modes, although the best-performing ensemble model remained dependent on the specific scenario. As shown in Figure 2, the highest macro F1 scores were achieved by Histogram Gradient Boosting for AHV (0.962), Histogram Gradient Boosting for ALO (0.902), and Extra Trees for FTO (0.952).

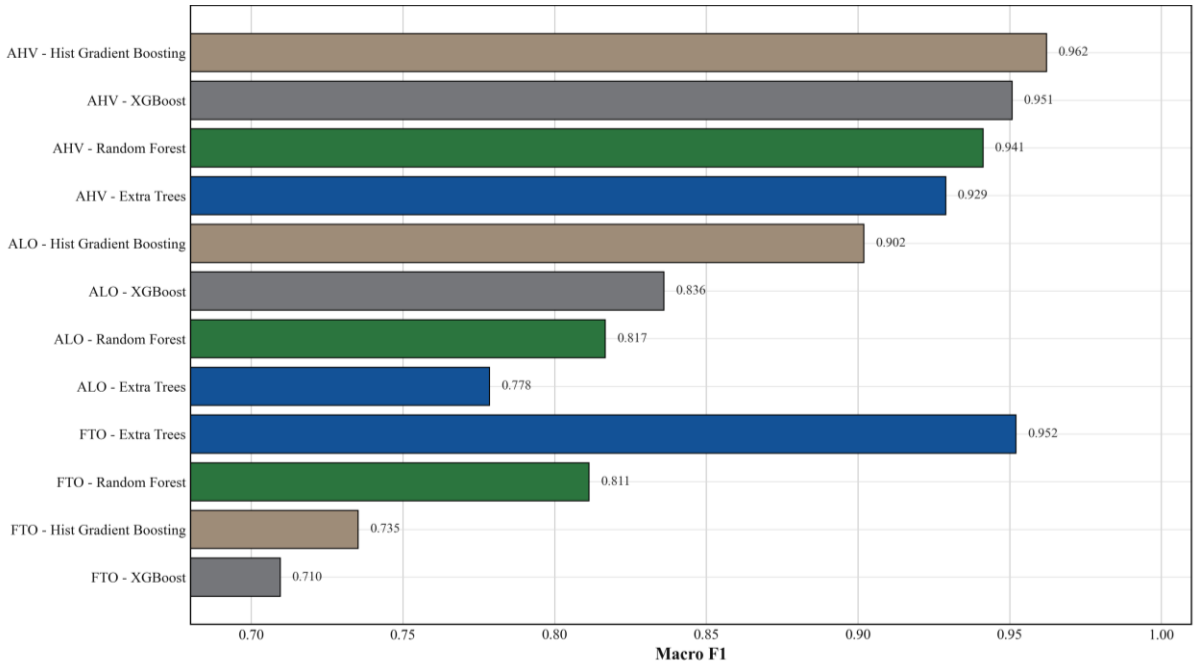

Figure 2: Ensemble Tree Model Comparison Across Failure Modes

The global SHAP results in Figure 3 show that the attribution patterns differ substantially across failure modes, both in the dominant feature families and in the magnitude of their SHAP values. AHV exhibits the largest attribution magnitudes, indicating that its classifier is driven by a set of features including pump motor current behavior, pump pressure, storage pressure response, and vaporizer pressure dynamics. ALO shows intermediate attribution magnitudes and a broader distribution of influential variables, with storage-pressure response, pump pressure behavior, vaporizer pressure change, and motor-current features contributing to the classification. FTO exhibits much smaller absolute SHAP magnitudes, with importance distributed across pump current statistics and storage-pressure dynamics. These results suggest that the degradation signatures associated with each failure mode differ in both strength and complexity. The findings indicate that health state diagnosis is supported by distinct physical response indicators that reflect different subsystems degradation.

The scale differences in Figure 3 provide additional insight into the nature of the degradation mechanisms. AHV appears to exhibit a sharply defined degradation signature dominated by a small number of highly influential variables, whereas ALO is characterized by a broader combination of

process-response indicators. In contrast, FTO is associated with lower-magnitude contributions distributed across multiple features, suggesting a more diffuse degradation pattern. These observations support the premise that pump degradation should be viewed as a multicomponent process involving interacting subsystems and that health state information can be extracted from operational response measurements that are more representative of underlying degradation processes.

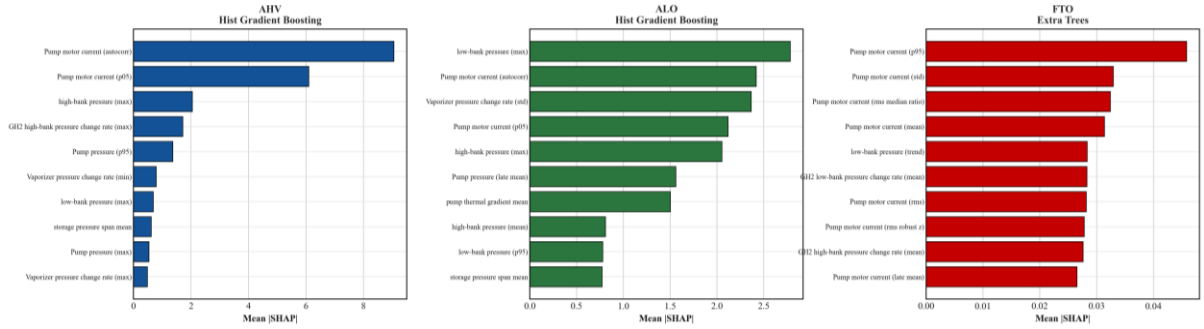

Figure 3: Top Global SHAP Features of Best Models

Representative local explanations are shown in Figure 4. These case-level attributions provide run-specific evidence for individual predictions and illustrate how different subsystems contribute to each health state classification. AHV and ALO show larger local attribution magnitudes, indicating that individual diagnostic decisions can be strongly influenced by a small number of dominant indicators. In contrast, the FTO example exhibits smaller contributions distributed across several variables, suggesting that the prediction is supported by a combination of multiple lower-magnitude degradation signals. For decision support purposes, these local explanations improve model transparency by linking diagnostic outcomes to operating conditions and subsystem behaviors.

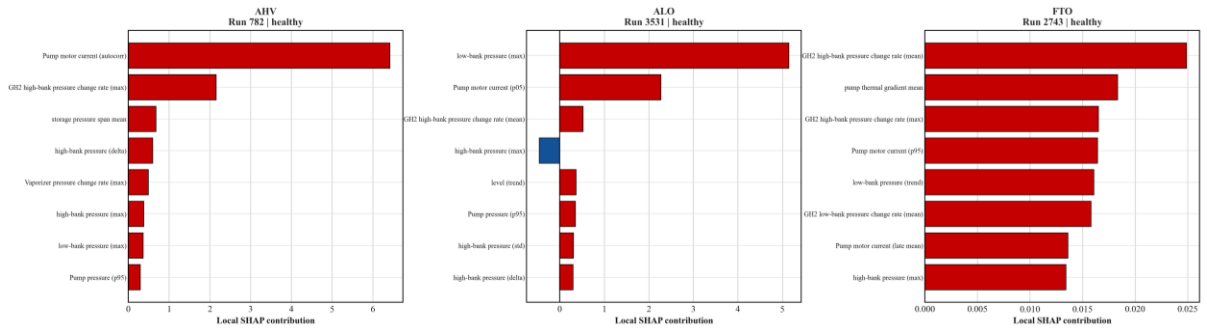

Figure 4: Representative Local SHAP Explanations

Overall, the results show that the different failure modes appear to have distinct operational signatures. AHV is characterized by a concentrated response pattern driven by pump electrical behavior, pressure, storage, and vaporizer pressure dynamics. ALO is more diffuse, relying on a broader set of process-response indicators across storage pressure, pump pressure, vaporizer change, and electrical features. FTO is the weakest in absolute attribution magnitude and appears to emerge from lower-magnitude contributions spread across pump current and storage-pressure dynamics. In practice, this means each failure mode should be monitored through different subsystems rather than with a one-size-fits-all indicator. For reliability engineering, the key implication is that health-state information can be extracted from process-response measurements alone, making diagnosis more robust to control action variability. The local SHAP explanations further support this by linking individual predictions to specific features, which is valuable for decision support. A challenge is represented by ALO, where the overlap between degraded stages suggests that additional subsystem-specific measurements or richer failure mode progression features may be needed to improve discrimination.

5. CONCLUSIONS

This study presented an explainable condition-based maintenance workflow for LH_2 pump systems that combines degradation-sensitive feature engineering, specific failure mode scenario health state classification, and explainable model outputs for run-level health state assessment. The results showed strong classification performance across health state for AHV, ALO, and FTO, while highlighting that the optimal ensemble learning strategy depends on the degradation characteristics of each failure mode. AHV and FTO exhibited clear health state separation, whereas ALO remained the most challenging case, particularly in distinguishing degraded and failed conditions.

The explainability analysis showed that predictions were driven by key operational system variables, with strong performance maintained even after removing direct control-action features. These findings indicate that the framework captures degradation-sensitive behavior directly from operational measurements. In the context of cryogenic hydrogen pumps, explainable diagnostics serve as an effective decision-support tool for effective CBM. Such capabilities support earlier detection of performance degradation, improve pump reliability, enhance station availability, and contribute to the safe operation of hydrogen fueling infrastructure. Future work will focus on improving health state separability, validating the framework across additional stations and operating conditions, and exploring a richer feature space to further improve the connection between model predictions and underlying degradation mechanisms.

Acknowledgements

This material is based upon work supported in part by the National Science Foundation under Grant No. 2045519. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work also received funding from the A. James Clark School of Engineering at the University of Maryland.

References

[1] S. Mosleh, C. Schaad, R. Yang, and K. M. Groth. *"A methodology for quantitative risk assessment of a high-capacity hydrogen fueling station with liquid hydrogen storage,"* International Journal of Hydrogen Energy, vol. 112, pp. 544–553, (2025).
[2] C. Schaad, S. Mosleh, R. Yang, and K. M. Groth. *"Quantitative Risk Assessment of hydrogen releases in a hydrogen fueling station with liquid hydrogen storage,"* International Journal of Hydrogen Energy, vol. 112, pp. 111–120, (2025).
[3] F. Zhang et al. *"Safety design of the piston rod seal structure for submerged high-pressure liquid hydrogen pump,"* Clean Energy, vol. 9, no. 1, pp. 108–122, (2025).
[4] K. M. Groth, L. M. Reising, V. Grabovetska, and A. Ruiz. *"Hydrogen Systems Risk and Reliability Challenges, Priorities, and Workshop Insights,"* Hydrogen Safety, vol. 19, no. 1, pp. 88–98, (2025).
[5] G. Petitpas and S. M. Aceves. *"Liquid hydrogen pump performance and durability testing through repeated cryogenic vessel filling to 700 bar,"* International Journal of Hydrogen Energy, vol. 43, no. 39, pp. 18403–18420, (2018).
[6] M. Nosaka. *"Cryogenic Tribology of High-Speed Bearings and Shaft Seals in Liquid Hydrogen,"* Tribology Online, vol. 6, no. 2, pp. 133–141, (2011).
[7] Z. Zhu, Y. Lei, G. Qi, Y. Chai, N. Mazur, Y. An, and X. Huang. "*A review of the application of deep learning in intelligent fault diagnosis of rotating machinery,"* Measurement, vol. 206, pp. 112346, (2023).
[8] R. Cassandro and K. M. Groth. *"When [Almost] Nothing Fails: The Data Paradox in Reliability Engineering,"* IEEE Reliability Magazine, (2026).
[9] R. Cassandro, C. Schaad, S. Marchetti, and K. M. Groth. *"Unsupervised Bayesian Health State Modeling for Predictive Reliability of Cryogenic Hydrogen Pumps,"* IEEE International Conference on Prognostics and Health Management (ICPHM 2026), pp. 1–6, (2026).

[10] R. Moradi and K. M. Groth. *"Modernizing Risk Assessment: A Systematic Integration of PRA and PHM Techniques,"* Reliability Engineering & System Safety, vol. 204, p. 107194, (2020).
[11] K. I. Matveev and J. W. Leachman. *"The Effect of Liquid Hydrogen Tank Size on Self-Pressurization and Constant-Pressure Venting,"* Hydrogen, vol. 4, no. 3, pp. 444–455, (2023).
[12] R. Cassandro, J. W. Rupe, and Z. S. Li. *"A hierarchical ensemble learning approach for cable network impairment diagnosis,"* IEEE Internet of Things Journal, vol. 11, no. 13, pp. 23068–23079, (2024).
[13] S. González, S. García, J. Del Ser, L. Rokach, and F. Herrera. *"A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities,"* Information Fusion, vol. 64, pp. 205–237, (2020).
[14] A. M. Elshewey, E. Selem, and A. H. Abed. *"Improved CKD classification based on explainable artificial intelligence with extra trees and BBFS,"* Scientific Reports, vol. 15, no. 1, pp. 17861, (2025).
[15] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al. *"LightGBM: A highly efficient gradient boosting decision tree,"* Advances in Neural Information Processing Systems, vol. 30, (2017).
[16] T. Chen and C. Guestrin. *"XGBoost: A scalable tree boosting system,"* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, (2016).
[17] S. M. Lundberg and S. I. Lee. *"A unified approach to interpreting model predictions,"* Advances in Neural Information Processing Systems, vol. 30, (2017).