© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. Accepted version; to appear in the 2026 IEEE International Conference on Big Data (BigData).

Auditing SHAP Explanations Against Known Failure Mechanisms in Predictive Maintenance

Koshik Debanath, Prakrit Uprety, Peiyu Li
Department of Electrical Engineering and Computer Science
South Dakota State University, Brookings, SD, 57007, USA
Email: {koshik.debanath, prakrit.uprety}@jacks.sdstate.edu, peiyu.li@sdstate.edu

***Abstract*—Machine learning models are widely used to predict equipment failures in industrial systems, but accurate predictions do not guarantee that models rely on features consistent with physical failure mechanisms. Although explainable artificial intelligence methods identify influential features, whether these explanations align with documented failure mechanisms is rarely tested statistically. To address this gap, we propose a mechanism-level statistical audit that evaluates whether feature attributions concentrate on mechanism-relevant features, distinguish among failure mechanisms, and reveal potentially misleading feature reliance. Three complementary statistics are evaluated against reference distributions generated by retraining models on shuffled labels. Experiments on the AI4I 2020 benchmark show evidence of attribution alignment for three of four documented failure mechanisms with engineered features, two of which remain aligned in every prespecified input and attribution configuration. Additional evaluations on laboratory bearing recordings and simulated industrial process data demonstrate the applicability of the framework across different settings. Controlled shortcut experiments show that models can maintain high predictive performance while relying on misleading features. A grouped statistic detects shortcuts spread over a few features, although weaker shortcuts may not be detected. Audit conclusions are also sensitive to the choice of explanation method and reference distribution. Overall, the proposed framework enables systematic statistical evaluation of explanation–mechanism alignment, while recognizing that attribution agreement does not establish physical correctness.**

***Index Terms*—Explainable AI, SHAP, predictive maintenance, feature attribution evaluation, physical failure mechanisms**

I. Introduction

Predictive maintenance is a maintenance strategy that uses machine-condition data to help anticipate faults and determine when maintenance should be performed. By identifying developing problems, it can help reduce unplanned downtime [1]. Machine-learning models support this process by analyzing equipment data to predict failures and maintenance needs. These models include decision trees, random forests, gradient-boosted trees, support vector methods, and neural networks [2], [3]. Complex ensembles and neural networks can be difficult to interpret, making prediction performance alone insufficient for understanding their behavior [4]. Understanding which sensor readings influence their predictions can help maintenance teams assess an alarm and decide what to investigate [5]. Interpretability approaches used in predictive maintenance include decision trees and local surrogate explanations such as Local Interpretable Model-agnostic Explanations (LIME) [2], as well as SHapley Additive exPlanations (SHAP) for interpreting sensor contributions to model predictions [5], [6]. Recent work has also examined how preprocessing and model complexity affect explanation quality in remaining-useful-life prediction [7].

High predictive performance can coexist with reliance on shortcuts or correlations unrelated to the intended task [8], [9]. In predictive maintenance, this raises the question of whether explanations emphasize the expected features for each failure type. SHAP assigns feature contributions to individual predictions [4], but these contributions do not by themselves establish the physical relevance of those inputs [10]. Evaluating feature attributions is difficult when a suitable ground-truth reference is unavailable [11]. Assessing mechanism alignment therefore requires information about which failure mechanism occurred and which features are expected to matter. Label-permutation methods provide a statistical reference for evaluating feature importance when the association between inputs and outcomes is disrupted [12]. The question is therefore not only which features receive attribution, but whether their attribution agrees with the recorded mechanism and exceeds that observed under this reference.

In this work, we investigate this question using the AI4I 2020 predictive maintenance dataset, a synthetic dataset containing machine operating measurements, an overall failure label, and separate failure-type labels [13]. Its recorded failure types and documented generation rules provide an external reference for evaluating mechanism alignment. Previous work on this dataset has assessed explanation quality using selected examples [14], interpreted SHAP feature importance [15], or quantitatively compared SHAP and LIME rankings [16]. Selected examples provide limited evidence across a failure type, while feature rankings and agreement between explainers do not by themselves establish agreement with its failure rule. This motivates our statistical audit of mechanism-level attribution on held-out observations, using models retrained on shuffled overall failure labels as a reference [12].

For the primary AI4I audit, we train a classifier to predict overall machine failure while withholding the individual failure-type labels from its inputs. We use SHAP to explain individual model predictions. Its feature contributions let us assess whether explanations emphasize the features specified by each failure rule. We compute SHAP values for tree ensembles using TreeSHAP [17]. To assess robust-

ness beyond tree models, we audit a multilayer perceptron (MLP) using KernelSHAP and DeepSHAP [4]. We group the SHAP explanations of individual model predictions by failure mechanism using the recorded failure-type labels. The audit assesses mechanism alignment by comparing held-out SHAP attributions with features identified as relevant by documented failure rules. Three statistics measure the attribution share on these features, how much larger that share is than on other failure types, and the largest excess share on a feature outside that set relative to other failures. Models retrained on shuffled overall failure labels provide a reference for assessing these patterns after disrupting the relationship between model inputs and failure outcomes. We first evaluate the audit on AI4I, then apply it to Case Western Reserve University (CWRU) laboratory bearing recordings [18] and simulated Tennessee Eastman process (TEP) data [19] to examine its applicability across different datasets and failure mechanisms. Supporting experiments assess sensitivity to models, explainers, and attribution references, and test the detection of planted shortcuts.

The paper makes the following contributions:

- We propose a statistical framework for auditing whether model explanations align with documented failure mechanisms. It combines mechanism-level attribution statistics, including a grouped outside-set statistic for distributed reliance, held-out explanations, a reference from models retrained on shuffled overall failure labels, a mechanism-specificity reference, and multiple-comparison correction.
- We investigate how alignment conclusions depend on statistical references, models, explainers, and feature definitions, and we give a prespecified protocol that reports each verdict with its sensitivity. Controlled shortcut experiments and calibration runs on independent generated data assess the audit's ability, false-flag rates, and limits in detecting misleading reliance that predictive performance alone can conceal.
- We demonstrate the framework's applicability across AI4I synthetic maintenance data, CWRU laboratory bearing recordings, and Tennessee Eastman process simulations. We assess robustness and scalability, and provide predictive benchmarks and artifacts for reproducing the experiments.

II. Related Work

Explainable AI methods, including LIME and SHAP, have been studied in predictive maintenance to help interpret machine-learning predictions [1], [5], [6]. Although small decision trees can provide interpretable rules, ensembles and neural networks often require additional explanation methods. Local explanations help examine individual alarms, while global summaries describe recurring patterns of feature importance. These provide different levels of insight into model behavior.

Maintenance studies apply explanation methods to both prognostics and fault diagnosis. Steurtewagen and Van den Poel [5] combine sensor and report data with XGBoost and use Shapley values to investigate machinery breakdowns in the petrochemical industry. Jalali et al. [3] compare LIME, SHAP, and attribute-wise explanations for models predicting the remaining useful life of aircraft engines using PHM08-CMAPSS sensor data. Gawde et al. [6] apply several explanation methods to rotating-machine fault diagnosis. Chen et al. [20] extend SHAP to the cyclic-spectral domain, relating carrier and modulation frequency contributions to bearing fault characteristics on CWRU data. In process monitoring, Gravanis et al. [21] compare Integrated Gradients and SHAP explanations of an LSTM classifier on TEP and examine their consistency with process knowledge. Together, these studies illustrate how explanations support prognostics and fault diagnosis, including assessment against domain knowledge.

On AI4I 2020, existing studies examine explanations from several perspectives. Matzka [14] assesses explanations against the dataset's failure rules using selected examples. Zhao et al. [15] examine SHAP feature importance, while Obeten and Jimoh [16] quantitatively compare SHAP and LIME rankings. Together, these studies assess physical plausibility, feature importance, and consistency between explanation methods.

The interpretation of feature attributions also depends on the assumptions made by explanation methods. SHAP assigns additive feature contributions relative to a baseline [4]. Covert et al. [22] explain that removal-based methods differ in how they represent features treated as missing when evaluating model predictions. Their framework highlights that assumptions about unavailable features are part of the explanation definition. Path-dependent TreeSHAP uses training-data coverage along tree paths, whereas interventional TreeSHAP averages over an explicit background dataset [17]. Feature dependence can also change how attribution is distributed, particularly when engineered inputs share information with their source measurements [23], [24]. For neural models, Integrated Gradients accumulates gradients along a path from a baseline to the explained observation [25], and baseline selection affects the resulting explanation [26].

Broader explanation-evaluation research examines several properties of attribution quality. XAI-Bench uses synthetic data for controlled evaluation [27], while OpenXAI evaluates properties including faithfulness, stability, and fairness [28]. ROAR removes highly attributed inputs, retrains the model, and measures changes in predictive performance [29]. Global sensitivity analysis based on the Hilbert–Schmidt Independence Criterion (HSIC) measures statistical dependence between inputs and outputs [30].

Statistical tests assess predictive performance and feature importance, while randomization checks examine how explanations respond to shuffled training labels or randomized model parameters. Label-permutation tests have been used to assess classifier performance [31], and outcome permutations have been used to calibrate feature-importance measures [12]. Adebayo et al. [32] showed that some visually convincing saliency methods remain insensitive to changes in model parameters or training labels. Recent work by Kim et al. [33] supplements global SHAP rankings for regression models with statistical summaries of direction, uncertainty, fidelity,

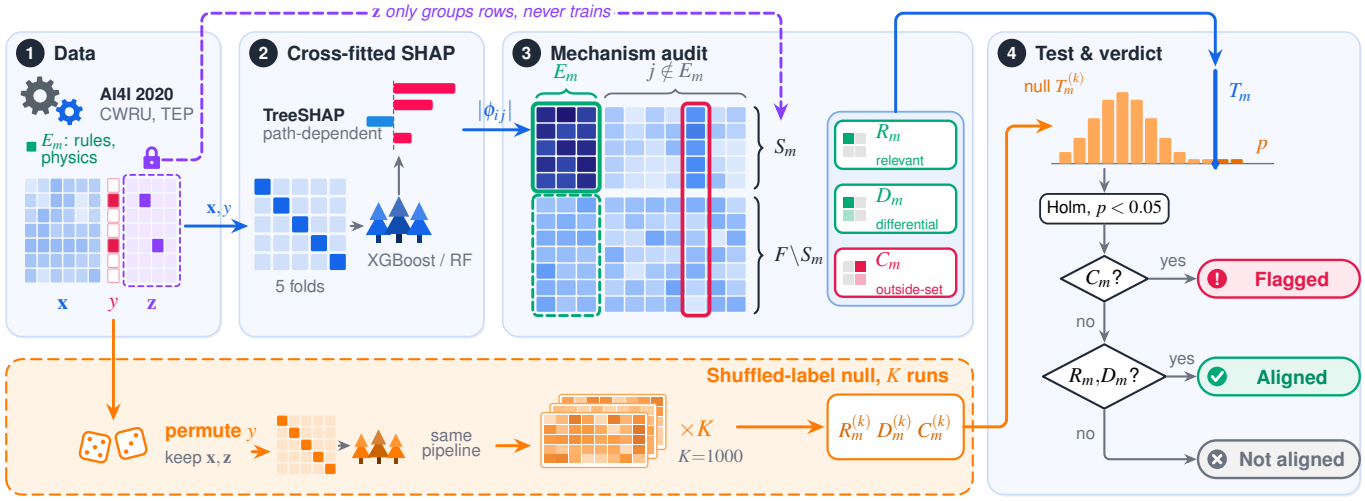


Fig. 1. Overview of the mechanism-level SHAP audit. (1) A classifier is trained using sensor features x and overall failure labels y, while mechanism indicators z are reserved for evaluation. (2) Cross-fitting produces held-out SHAP explanations for each observation. (3) For each mechanism m, absolute attributions are compared between mechanism-positive observations S_m and other failures $F \setminus S_m$. The expected features E_m (green) and remaining features are used to calculate the relevant share R_m, differential score D_m, and outside-set excess C_m (red). Cell values are illustrative. (4) Each statistic is compared with a reference distribution from K models retrained on shuffled failure labels, keeping x and z unchanged. After Holm correction, a significant C_m triggers a flag; otherwise, significant R_m and D_m indicate alignment. Remaining cases are classified as not aligned.

and stability. Beyond these statistical assessments, shortcut-learning research shows that accurate predictions can rely on unintended cues, motivating closer examination of which features models use [8], [9].

Existing research provides complementary ways to interpret predictions and evaluate feature attributions. However, feature rankings, agreement between explainers, and faithfulness to model behavior do not by themselves establish alignment with documented failure mechanisms. Our work complements these approaches through a statistical audit of mechanism-level alignment and attribution to features outside those identified as relevant by documented failure rules.

III. Methodology

A. Problem Definition

Let $\mathcal{D} = \{(x_i, y_i, z_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^d$ contains the input features, $y_i \in \{0,1\}$ indicates overall failure, and $z_i \in \{0,1\}^M$ records M failure mechanisms. A classifier f is trained using only (x_i, y_i); the mechanism indicators are reserved for evaluating its explanations. For each mechanism m, domain knowledge defines a set E_m of expected features. Let $S_m = \{i : z_{i,m} = 1\}$ denote mechanism-positive observations, $F = \{i : y_i = 1\}$ all labeled failures, and $\phi_j(x_i)$ the SHAP attribution of feature j for observation i.

Our goal is to test whether explanations emphasize the expected features when a particular failure mechanism occurs, distinguish that mechanism from other failures, and reveal unusual reliance on features outside the expected set. The audit measures agreement with a predefined mechanism description, not whether the model has discovered the physical cause of failure.

B. Overview of the Proposed Audit

Figure 1 illustrates the four stages of the audit: training without mechanism labels, cross-fitted SHAP explanations, mechanism-level statistics, and comparison with models retrained on shuffled labels. A significant outside-set statistic produces a flagged verdict. Otherwise, a mechanism is aligned when both alignment statistics are significant and not aligned when they are not. The following subsections define each step.

C. Audit Statistics

We use absolute SHAP values to measure how strongly each feature contributes to the model's predictions, regardless of whether the contribution increases or decreases the predicted failure score. For feature j and observation set A, we first calculate its mean absolute attribution and then normalize it by the total attribution across all features:

$$I_j(A) = \frac{1}{|A|} \sum_{i \in A} |\phi_j(x_i)|, \qquad s_j(A) = \frac{I_j(A)}{\sum_{k=1}^{d} I_k(A)}. \quad (1)$$

Here, $I_j(A)$ measures the average attribution of feature j, while $s_j(A)$ represents its share of the total attribution. For mechanism m, let $r_m(A) = \sum_{j \in E_m} s_j(A)$ denote the combined share assigned to its expected features. Using these quantities, we define three complementary statistics:

$$R_m = r_m(S_m), \quad (2)$$

$$D_m = r_m(S_m) - r_m(F \setminus S_m), \quad (3)$$

$$C_m = \max_{j \notin E_m} \left[s_j(S_m) - s_j(F \setminus S_m) \right]. \quad (4)$$

The *relevant share* R_m measures how much attribution is assigned to the expected features when mechanism m occurs.

However, these features may also be important for other failures, so a large R_m alone does not establish mechanism-specific alignment. The *differential score* D_m addresses this by comparing the expected-feature share on mechanism-positive observations with that on other failures. Finally, the *outside-set excess* C_m measures the largest increase in attribution share among features outside E_m, allowing the audit to identify unexpected feature reliance that the first two statistics may overlook.

Because C_m takes a maximum, reliance spread over several outside features can escape it. We therefore also report a grouped outside-set excess,

$$C_m^{(3)} = \sum_{j \in T_m} \max[0,\ s_j(S_m) - s_j(F \setminus S_m)]\,, \tag{5}$$

where T_m holds the three outside features with the largest excess. We fixed $k = 3$ before running the revision experiments. $C_m^{(3)}$ is tested in the same way as C_m. It is reported next to each verdict as a secondary flag and does not change the primary decision rule.

Together, the statistics examine whether explanations emphasize the expected features, distinguish the mechanism from other failures, and reveal unexpected attribution. Their statistical significance is determined using the shuffled-label reference described in Section III-D. A significant outside-set excess requires further investigation because it may reflect shortcut reliance, an incomplete expected feature set, or a downstream effect rather than an incorrect explanation.

D. Statistical Inference

a) Shuffled-label reference: A large attribution share does not necessarily indicate mechanism alignment, as unusual sensor values may receive substantial attribution even when a model has learned no meaningful relationship with failure labels. To account for this, we randomly shuffle the overall failure labels y while keeping the sensor measurements x, mechanism indicators z, and expected feature sets E_m unchanged. For each permutation, we repeat the training and explanation pipeline using the selected model settings and calculate the three audit statistics. The resulting distributions provide a reference for testing whether the observed attribution patterns exceed those obtained after disrupting the input–label relationship.

b) Significance testing: For each statistic $T \in \{R_m, D_m, C_m\}$, we compare its observed value with the corresponding values from K shuffled-label pipelines. The one-sided Monte Carlo p-value is calculated as

$$p = \frac{1 + \sum_{k=1}^{K} \mathbb{I}(T^{(k)} \geq T)}{K + 1}. \tag{6}$$

The primary audit uses $K = 1{,}000$ permutations. To account for testing multiple mechanisms, we apply Holm's correction [34] across the four physical mechanisms separately for each statistic, using a significance level of $\alpha = 0.05$. With K permutations and M mechanisms, the smallest attainable adjusted p-value is $M/(K+1)$, which is 0.004 for the primary audit. Equal p-values at this floor therefore do not indicate equal evidence, so we also report standardized effects.

c) Mechanism-specificity reference: The shuffled-label reference asks whether a model has learned a relationship with failure. It does not ask whether attribution is specific to a particular mechanism. We therefore add two references that keep the fitted model and its held-out SHAP values fixed. The first permutes the mechanism indicator vectors z_i among labeled failures ($B = 10{,}000$ draws), which preserves the size of each S_m. The second enumerates every expected set of size $|E_m|$ that can be formed from the model inputs. We compare R_m and D_m with both references. These references cannot replace the retrained one, because a model trained on shuffled labels can still show mechanism-specific attribution when mechanism observations have unusual sensor values (Section V-A).

d) Audit Verdicts: Following the decision process in Fig. 1, we first examine the outside-set statistic C_m. A significant result produces a ***flagged*** verdict, indicating unexpected attribution that requires further investigation. Otherwise, a mechanism is considered *aligned* only when both R_m and D_m are significant. If these conditions are not satisfied, the mechanism is classified as *not aligned*, indicating insufficient statistical evidence of agreement under the tested configuration. Importantly, alignment does not establish physical causality, and the absence of an outside-set flag does not guarantee that the model is free of shortcuts.

To supplement the significance tests, we report standardized differences from the null distribution, $z = (T - \bar{T}_0)/s_0$, where $\bar{T}_0$ and s_0 are the null mean and standard deviation. For the primary AI4I audit, we also report 95% bootstrap intervals for R_m and D_m using 5,000 resamples. These intervals describe uncertainty conditional on the fitted models, while the shuffled-label distributions provide the reference for statistical testing. Implementation details and pseudocode are available in the reproduction artifacts [35].

IV. Experimental Design

We use AI4I 2020 as the primary benchmark because its documented failure rules provide a reference for evaluating mechanism alignment. We examine the sensitivity of the audit to different models, explanation methods, and feature sets, and introduce controlled shortcuts to test its ability to detect misleading feature reliance. We then apply the audit to CWRU bearing recordings and Tennessee Eastman process simulations, with dataset-specific characteristics and evaluation procedures described in Section IV-E.

A. Dataset and Failure Mechanisms

AI4I 2020 [14] is a synthetic predictive maintenance dataset with 10,000 observations, including 339 labeled failures. Its inputs include product quality, air and process temperatures, rotational speed, torque, and tool wear. The dataset records five mechanisms: tool wear failure (TWF), heat dissipation failure (HDF), power failure (PWF), overstrain failure (OSF), and

TABLE I
AI4I 2020 FAILURE MECHANISMS [14], THEIR TRIGGER CONDITIONS, EXPECTED FEATURES E_m, AND MECHANISM-POSITIVE OBSERVATIONS n. L/M/H DENOTE PRODUCT QUALITY TYPES.

Mech.	Trigger	E_m	n
TWF	Tool wear at a random point in 200–240 min	Tool wear	46
HDF	(Process − air temp.) < 8.6 K and speed < 1380 rpm	Both temperatures, their difference, speed	115
PWF	Torque × speed (rad/s) outside [3500, 9000] W	Torque, speed, power	95
OSF	Tool wear × torque above 11,000/12,000/ 13,000 min·Nm for L/M/H	Tool wear, torque, overstrain, its ratio	98
RNF	0.1% chance, independent of the process	–	19

random failure (RNF). Their documented rules and expected features are summarized in Table I.

The released labels contain inconsistencies: 24 observations have multiple mechanism flags, nine labeled failures have none, and 18 of the 19 RNF observations are labeled as non-failures. We retain the released labels for the primary experiments and examine corrected labels and single-mechanism observations separately. The audit covers the four mechanisms with documented physical rules; RNF has no expected feature set and is examined descriptively.

B. Feature Construction

Several failure rules combine sensor measurements. We therefore construct mechanical power $P = \tau(2\pi n/60)$, temperature difference $\Delta T = T_{\text{process}} - T_{\text{air}}$, overstrain $O = w\tau$, and overstrain ratio $Q = O/L(t)$. Here, τ, n, and w denote torque, rotational speed, and tool wear, while $L(t)$ is the type-specific overstrain limit of 11,000, 12,000, or 13,000 for L, M, and H. Each expected set E_m contains the measurements and derived quantities used to express its failure rule. These features use no failure labels but encode known benchmark rules; we therefore also audit the original features alone to assess their influence on the results.

C. Model Selection and Held-Out Explanations

For AI4I, we use stratified training, validation, and test splits of 70%, 20%, and 10%. Preprocessing includes scaling and one-hot encoding fitted on the training data. Borderline-SMOTE [36] is selected among imbalance treatments using validation F1. XGBoost [37] hyperparameters are selected through 40-iteration randomized search with five-fold cross-validation, with oversampling performed inside each training fold. The decision threshold is chosen on the validation set, and the reserved test set measures predictive performance.

For the audit, these choices are fixed and the complete pipeline is refitted using five-fold stratified cross-fitting over all observations. Each observation is explained by a model that did not train on it. The primary audit uses path-dependent TreeSHAP [17]; shuffled-label pipelines repeat preprocessing, oversampling, fitting, cross-fitting, and explanation generation. Model selection is not nested within the audit folds, so we separately examine the effect of re-tuning on each permutation. Other classifiers and explanation methods are included in the sensitivity experiments, with full settings reported in the reproduction artifacts [35].

D. Prespecified Audit Protocol

Our sensitivity results show that one fitted model can receive different verdicts under different explanation configurations. To keep a practitioner from choosing whichever configuration gives the preferred conclusion, we fix the following steps before any attributions are computed. (i) *Expected sets.* E_m contains every measured input named in the documented rule for m, together with any derived quantity in which the rule's threshold is stated. Inputs that are neither named nor used to state the threshold are excluded. (ii) *Inputs.* The audit is run on the raw measurements and on the rule-encoding inputs, because derived features can make a rule easier to attribute. (iii) *Reference population.* For questions about abnormal operation, we recommend an interventional reference drawn from normal operation. Path-dependent TreeSHAP, whose reference is the (possibly oversampled) training distribution, is reported alongside it. (iv) *Reporting.* Each mechanism is audited on the resulting 2×2 set (raw or engineered inputs, path-dependent or normal-operation reference). A verdict is called *stable* only when all four configurations agree, and *configuration-dependent* otherwise. We report it together with agreement across the other explainers, the mechanism-specificity references, and the grouped statistic $C_m^{(3)}$. On AI4I, we applied this protocol after the primary analysis, so it summarizes our results rather than validating the protocol prospectively.

E. Evaluation Settings Across Datasets

a) AI4I variants and shortcuts: We vary feature construction, imbalance treatment, model family, explanation method, attribution reference, and null-model training, comparing each configuration with its own shuffled-label reference. To introduce a shortcut, we add $s_i = z_{i,\text{HDF}} + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. This noisy copy of the HDF indicator has no physical role and is excluded from E_{HDF}. We vary σ from 2 to 0.1 and examine shortcuts distributed over several sensors. Scaling experiments generate up to 3×10^5 observations from the documented AI4I process. Variants use $K = 200$ permutations unless stated otherwise; the shortcut sweep uses $K = 500$.

b) CWRU bearing recordings: We use 36 laboratory vibration recordings [18] covering inner-race, ball, and outer-race faults. A random forest classifies fault location from 19 features extracted from 4,096-sample windows. Expected features are spectral bands around each fault's characteristic defect frequency and second harmonic, with a primary bandwidth of ± 0.15 shaft orders. Four-fold group cross-fitting holds out complete recordings, and the null shuffles labels across recordings ($K = 1{,}000$). SHAP values explain each observation's recorded fault class, with the other fault classes forming its comparison group. We also introduce a

fault-dependent acquisition-gain shortcut that affects vibration amplitude.

c) Tennessee Eastman process: We use 400 simulated runs [19] containing reactor-cooling (IDV4) and condenser-cooling (IDV5) disturbances. Each run contributes 16 observations before and 16 after fault activation, using 53 process variables; the expected sets contain three variables associated with each disturbed cooling loop. XGBoost uses five-fold group cross-fitting over complete runs. The primary null shuffles fault-activity labels within runs ($K = 200$), while a second null rotates the ordered labels to preserve a contiguous fault-active block. We also examine 30-minute temporal features and an LSTM explained with Integrated Gradients. Neither temporal permutation scheme alone establishes exchangeability of the labels.

V. Experimental Results and Discussion

A. Main Audit Results

a) Predictive performance: The tuned XGBoost classifier achieves a test F1-score of 0.89, with precision 1.00 and recall 0.79. Because the test set contains only 34 failures, performance differences from other classifiers are not conclusive. Our main interest is whether the model's explanations agree with the documented failure mechanisms.

Table II shows that HDF, PWF, and OSF pass both alignment tests without significant outside-set flags (adjusted $p = 0.004$, $z = 3.4$–8.4). Their explanations also emphasize different features: speed and temperature difference for HDF, power and torque for PWF, and the overstrain ratio for OSF. This mechanism-specific pattern would be obscured by a global feature-importance ranking. HDF has the smallest excess relevant share, partly because its expected features already receive substantial attribution in shuffled-label models (null mean 0.56).

TWF does not pass the primary audit ($p_R = 0.201$, $p_D = 0.077$). Although tool wear is its most attributed feature, it receives only 19% of attribution, compared with 15% under the null. The classifier also detects few TWF observations (held-out recall 0.07), consistent with the partly random trigger in its failure rule. Thus, the primary audit supports agreement for three mechanisms but provides insufficient evidence for TWF under the tested configuration. We examine this difference further in the sensitivity analysis.

b) Raw inputs: Part of this alignment comes from the input design. Power, temperature difference, and the overstrain ratio encode the generation rules and belong to the expected sets. The lower block of Table II therefore repeats the audit with the five raw measurements only. PWF remains aligned, and OSF remains aligned with weaker specificity ($D_m = 0.160$, $z = 1.9$, $p = 0.048$). HDF passes the differential test ($z = 6.7$) but not the relevant-share test ($p = 0.190$), because shuffled-label models already place 63% of HDF attribution on air temperature, process temperature, and speed. TWF remains not aligned. With engineered inputs, the audit thus tests whether attribution follows known rules expressed in their natural units. The raw audit is closer to what a practitioner observes without such rules, and we report both in Table IV.

c) Mechanism specificity: Both model-fixed references give strong evidence of specificity. Permuting mechanism labels among failures gives $z = 13.5$–15.7 for D_m (adjusted $p = 0.0004$, the floor for $B = 10{,}000$). Among all expected sets of equal size, the documented set ranks 1st of 12 (TWF), 4th of 495 (HDF), 15th of 220 (PWF), and 15th of 495 (OSF) by D_m. PWF and OSF rank below some random sets mainly because their expected sets share speed, torque, or tool wear with other mechanisms. A model trained on shuffled labels also passes the label-permutation reference ($z = 6.7$–9.4), and it ranks 2nd of 220 for PWF among random sets. Mechanism observations occupy distinct regions of the input space, so attributions differ between mechanisms even when the model has learned nothing. Specificity therefore needs to be read together with the retrained reference. It does not replace it.

B. Ablation and Sensitivity Analysis

a) Audit Component Ablations: We examine the importance of three audit components: held-out explanations, a retrained shuffled-label reference, and the outside-set statistic. A model trained on shuffled labels fails the proposed alignment tests, but a naive comparison with random failure observations makes all four mechanisms appear significant ($p = 0.002$). Similarly, using in-sample explanations produces a TWF alignment verdict ($p = 0.004$) that does not hold under held-out evaluation, with its relevant share increasing from 0.19 to 0.35. Finally, the relevant-share test alone misses a strong planted shortcut in the large-sample experiment, whereas the outside-set statistic detects it. These results demonstrate how the three components help prevent misleading conclusions about mechanism alignment.

b) Robustness of the verdicts: Table III summarizes how the verdicts change across configurations. Random seeds, single-mechanism observations, corrected labels, and re-tuned shuffled-label models largely preserve the primary conclusions. However, removing engineered features causes HDF to lose alignment ($p_R = 0.19$; Table II), and replacing Borderline-SMOTE with SMOTE or ADASYN has a similar effect ($p_R = 0.109$). PWF and OSF are more stable under these changes. Supporting checks using attribution direction, feature removal, and statistical dependence are consistent with the main HDF, PWF, and OSF findings but do not independently establish physical causality.

c) Dependence on models and explanation methods: HDF, PWF, and OSF remain aligned under a random forest and under interventional TreeSHAP, unless the background contains only failures. Neural-model explanations agree less. On the MLP, Integrated Gradients aligns HDF and OSF, DeepSHAP aligns only HDF, and KernelSHAP aligns none. These comparisons involve differences in models and attribution definitions, so they cannot isolate the effect of the explainer alone. They do show that mechanism alignment depends on the explanation configuration, which motivates reporting each verdict with its sensitivity (Table IV).

TABLE II
AUDIT USING HELD-OUT SHAP VALUES (PATH-DEPENDENT TREESHAP, SEED 42, $K = 1{,}000$) WITH ENGINEERED INPUTS (PRIMARY) AND WITH RAW INPUTS ONLY.

	n	Top feature	R_m [95% CI]	Null	z	p	D_m [95% CI]	Null	z	p	C_m	Null	p	Verdict
Engineered inputs (primary)														
TWF	46	tool wear	0.193 [0.17, 0.22]	0.152	0.9	0.201$^{\times}$	0.129 [0.10, 0.16]	0.070	1.5	0.077$^{\times}$	0.025	0.025	1.000	not aligned
HDF	115	speed	0.711 [0.69, 0.73]	0.557	3.4	0.004	0.431 [0.41, 0.45]	0.049	8.4	0.004	0.001	0.015	1.000	aligned
PWF	95	power	0.696 [0.67, 0.72]	0.347	7.6	0.004	0.330 [0.30, 0.36]	0.088	5.7	0.004	0.001	0.011	1.000	aligned
OSF	98	overstrain ratio	0.672 [0.66, 0.69]	0.345	5.6	0.004	0.424 [0.40, 0.44]	0.067	7.0	0.004	-0.001	0.019	1.000	aligned
RNF	19	speed	0.186	0.240	–	0.985	–	–	–	–	–	–	–	–
Raw inputs only														
TWF	46	tool wear	0.272 [0.24, 0.30]	0.249	0.4	0.345$^{\times}$	0.075 [0.04, 0.11]	0.092	-0.4	0.616$^{\times}$	0.024	0.022	1.000	not aligned
HDF	115	air temp.	0.685 [0.66, 0.71]	0.626	1.3	0.190$^{\times}$	0.348 [0.32, 0.38]	0.073	6.7	0.004	0.000	0.009	1.000	not aligned
PWF	95	torque	0.642 [0.62, 0.67]	0.396	4.9	0.004	0.239 [0.21, 0.27]	0.099	3.0	0.004	-0.001	0.008	1.000	aligned
OSF	98	tool wear	0.624 [0.60, 0.64]	0.390	4.0	0.004	0.160 [0.12, 0.19]	0.065	1.9	0.048	0.013	0.012	1.000	aligned

Note: Null: mean over shuffled-label pipelines; z: (observed − null mean)/null s.d.; brackets: 95% bootstrap intervals conditional on fitted models; p: Holm-adjusted over four physical mechanisms; $^{\times}$: non-significant alignment test.

TABLE III
VERDICTS FOR SELECTED AUDIT VARIANTS: ✓ ALIGNED, × NOT ALIGNED, AND ! FLAGGED. FULL ADJUSTED p-VALUES ARE PROVIDED IN THE REPRODUCTION ARTIFACTS.

Variant	K	TWF	HDF	PWF	OSF
Primary and null					
Primary audit	1000	×	✓	✓	✓
Seeds 1–4 (largest p)	200	×	✓	✓	✓
Null re-tuned per permutation	500	×	✓	✓	✓
Data and expected sets					
Raw features only	1000	×	×	✓	✓
Label = OR of flags	200	×	✓	✓	✓
E_m without top trigger	200	×	!	!	!
Imbalance treatment					
Class weights	200	×	✓	✓	✓
Focal loss	200	✓	✓	✓	✓
SMOTE	200	×	×	✓	✓
Attribution (XGBoost)					
Interventional, training bg.	200	✓	✓	✓	✓
oversampled bg.	200	✓	✓	✓	✓
failure-only bg.	200	!	×	×	✓
Pairwise, nearest normal	200	✓	✓	✓	✓
Conditional group Shapley	200	✓	✓	✓	✓
LIME	200	✓	✓	×	✓
Other models					
Random forest, path TreeSHAP	200	×	✓	✓	✓
MLP, KernelSHAP	200	×	×	×	×
MLP, Integrated Gradients	200	×	✓	×	✓
MLP, SmoothGrad	200	!	✓	✓	✓
Sanity check					
Permuted labels	200	×	×	×	×

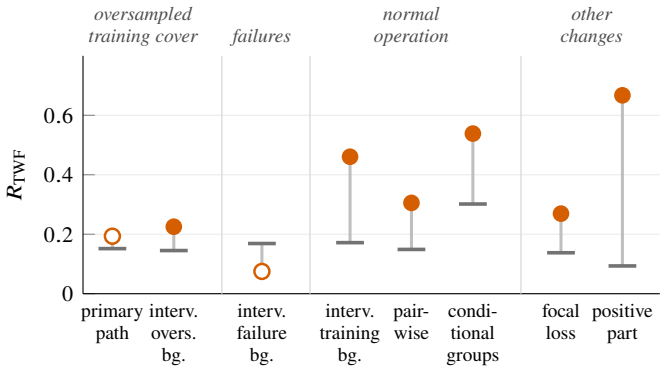


Fig. 2. Relevant share for TWF under different attribution references. Filled markers indicate alignment; bars show each configuration's shuffled-label null mean.

d) The TWF attribution reference: TWF provides the clearest example of attribution-reference dependence (Fig. 2). It is not aligned under path-dependent TreeSHAP, which measures attributions against the oversampled training data. Against that reference, ordinary sensor values count as evidence against failure and shrink tool wear's share of absolute attribution. Measured against normal operation, the same model passes and tool wear's share more than doubles. Oversampling alone is not the full story: an oversampled interventional background also passes, although its share of 0.23 stays close to the path-dependent 0.19. A failure-only reference removes the alignment. Additional checks suggest that feature correlation alone does not explain the difference. The evidence therefore points to the attribution reference as an important factor, without establishing it as the sole cause.

e) Verdicts with sensitivity: Table IV applies the pre-specified protocol. PWF and OSF are aligned in all four configurations and are reported as stable. TWF is aligned only against the normal-operation reference, with raw and with engineered inputs. HDF is aligned in every configuration except raw inputs with path-dependent attribution. Both are therefore reported as configuration-dependent, and the table states which configuration changes each verdict. Across the 13 explanation configurations, HDF and OSF are aligned in 11, PWF in 8, and TWF in 6. MLP KernelSHAP aligns no mechanism. The MLP's shuffled-label models already place much attribution on the expected features (R_{PWF} null mean 0.60–0.63, compared with 0.35 for XGBoost), so its reference is harder to exceed. Its predictive performance is also lower (test F1 0.76). We therefore recommend reporting cross-explainer agreement as part of a verdict rather than relying on a single explainer.

TABLE IV
AI4I VERDICTS REPORTED WITH THEIR PRESPECIFIED SENSITIVITY SET (SECTION IV-D).

	Prespecified set				Specificity D_m				
	Eng.		Raw		Expl.	Label	Set	$C_m^{(3)}$	
Mech.	path	norm.	path	norm.	agree	perm. z	rank	p	Reported
TWF	×	✓	×	✓	6/13	13.5	1/12	1.000	dependent
HDF	✓	✓	×	✓	11/13	15.7	4/495	1.000	dependent
PWF	✓	✓	✓	✓	8/13	13.6	15/220	1.000	stable
OSF	✓	✓	✓	✓	11/13	15.6	15/495	1.000	stable

Note: path: path-dependent TreeSHAP; norm.: interventional TreeSHAP with a training background (97% normal operation). Expl. agree: aligned verdicts among 13 explanation configurations (XGBoost with path-dependent, four interventional backgrounds, pairwise, conditional group Shapley, and LIME; random forest; MLP with KernelSHAP, Integrated Gradients, DeepSHAP, and SmoothGrad). Label perm. z: D_m against mechanism labels permuted among failures. Set rank: rank of D_m among all equal-size expected sets. $C_m^{(3)}$: Holm-adjusted p of the grouped outside-set test (engineered, path). Reported: stable if all four prespecified configurations agree, otherwise configuration-dependent.

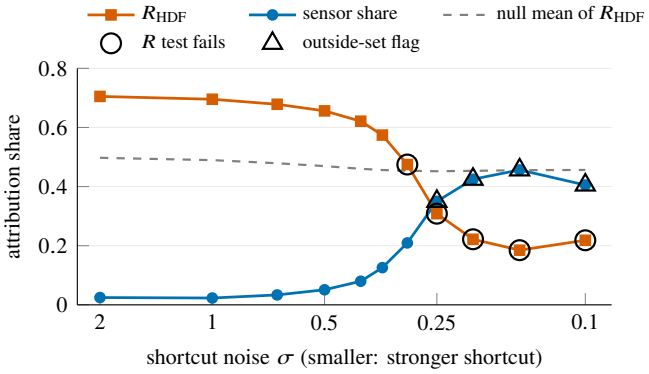


Fig. 3. Effect of shortcut strength on HDF attribution. As the planted sensor becomes more informative, its attribution share rises and mechanism-relevant attribution falls. Circles indicate failed alignment tests; triangles indicate outside-set flags.

C. Shortcut Detection

The planted-shortcut experiments test whether the audit can reveal misleading reliance without a loss in predictive performance. As the shortcut becomes stronger, its attribution increases while HDF's relevant share decreases (Fig. 3). Weak shortcuts ($\sigma \geq 0.35$) leave HDF aligned in this sweep, although $\sigma = 0.35$ cost HDF its alignment in two of three generated datasets of the same size (Table VII). At $\sigma = 0.30$, the mechanism loses alignment but no outside feature is flagged. Stronger shortcuts ($\sigma \leq 0.25$) are flagged by the outside-set test. Test F1 remains between 0.84 and 0.92, showing that predictive performance alone does not reveal this reliance.

Detection becomes harder when the shortcut is distributed across several features (Table V). With three or five sensors, no single sensor exceeds 25% of HDF attribution, and C_m flags only the strongest three-sensor case. The grouped statistic $C_m^{(3)}$ flags three sensors from $\sigma = 0.35$ (38% combined share; $p = 0.024$, compared with 0.90 for C_m) and five sensors from $\sigma = 0.30$ (53%; $p = 0.024$, compared with 0.77). For one sensor the two statistics coincide. Neither statistic flags any configuration with $\sigma \geq 0.5$, any outside feature of TWF, PWF, or OSF in these runs, or any mechanism in the unmodified AI4I audit ($p = 1.0$ with engineered and raw inputs). The grouped statistic does, however, produce more false flags on independent generated data, as reported with Table VII. At intermediate strengths, such as $\sigma = 0.40$ with 32–41% combined share, neither statistic flags, but R_{HDF} fails and HDF is reported as not aligned.

The experiments therefore establish both the value and the detection limits of the audit. Strong shortcuts can be identified despite high predictive accuracy, and the grouped statistic extends detection to shortcuts spread over a few features. An absent flag, however, is not evidence that no shortcut exists. It means only that no outside group of up to three features exceeds its null reference. Weaker shortcuts, or shortcuts spread over more features, may only reduce alignment or remain undetected.

TABLE V
DISTRIBUTED SHORTCUTS ($K = 500$): COMBINED HDF ATTRIBUTION SHARE OF 1, 3, OR 5 PLANTED SENSORS BY SENSOR NOISE σ. C: FLAGGED BY C_m; G: FLAGGED BY GROUPED $C_m^{(3)}$; ×: HDF NOT ALIGNED WITHOUT A FLAG.

Sensors	1	0.67	0.5	0.4	0.35	0.3	0.25
1	0.02	0.03	0.05	0.08	0.13	0.21 ×	0.35 CG
3	0.07	0.11	0.16	0.32 ×	0.38 G	0.48 G	0.54 CG
5	0.10	0.14	0.23	0.41 ×	0.47 ×	0.53 G	0.56 G

TABLE VI
AUDIT ON CWRU BEARING DATA ($K = 1{,}000$) AND SIMULATED TENNESSEE EASTMAN PROCESS DATA ($K = 200$).

Data	Fault	R_m	p	D_m	p	C_m	p	Verdict
CWRU	Inner	0.440	0.004	0.190	0.030	0.065	1.000	aligned
	Ball	0.133	0.773$^{\times}$	0.036	0.211$^{\times}$	0.051	1.000	not aligned
	Outer	0.392	0.003	0.131	0.027	0.042	1.000	aligned
+ gain	Inner	0.394	0.018	0.168	0.054$^{\times}$	0.063	0.989	not aligned
	Ball	0.105	0.948$^{\times}$	0.029	0.547$^{\times}$	0.020	0.989	not aligned
	Outer	0.269	0.018	0.050	0.547$^{\times}$	0.165	0.006†	flagged
TEP	IDV4	0.747	0.010	0.568	0.010	0.000	1.000	aligned
	IDV5	0.221	0.010	0.164	0.010	0.026	0.010†	flagged

Note: p: Holm-adjusted; $^{\times}$: non-significant alignment test; †: significant outside-set test.

D. Generalization and Scalability

a) Application to bearing and process data: Table VI reports the audit on two additional datasets. On CWRU, the classifier identifies all 36 held-out recordings correctly (window accuracy 0.99), and inner- and outer-race faults align with their expected defect-frequency features. Ball faults are not aligned under the primary frequency bands but become aligned with wider bands, showing the importance of the expected feature definition. When a fault-dependent acquisition-gain shortcut is introduced, classification remains perfect, yet

TABLE VII
VERDICTS ON GENERATED DATA OF INCREASING SIZE ($K = 200$): NUMBERS OF RUNS ALIGNED (A), NOT ALIGNED (N), AND FLAGGED (F). SHARE DENOTES PLANTED-SENSOR ATTRIBUTION; MIN. DENOTES MINUTES PER FULL AUDIT ON EIGHT CORES.

				HDF verdicts			TWF verdicts			
n	σ	Runs	Share	A	N	F	A	N	F	Min.
10,000	none	3	–	3	0	0	0	3	0	3
	0.35	3	0.158	1	2	0	0	3	0	3
	0.25	3	0.390	0	0	3	0	3	0	3
30,000	none	3	–	3	0	0	0	2	1	7
	0.35	3	0.128	3	0	0	0	3	0	6
	0.25	3	0.346	0	0	3	0	2	1	7
100,000	none	3	–	3	0	0	0	1	2	17
	0.35	3	0.100	3	0	0	0	2	1	19
	0.25	3	0.315	0	0	3	0	2	1	19
300,000	none	1	–	1	0	0	0	1	0	52
	0.35	1	0.063	1	0	0	0	1	0	59
	0.25	1	0.271	0	0	1	0	0	1	61

attribution shifts toward log-RMS and the outside-set test flags the change ($C_m = 0.165$, $p = 0.006$).

On Tennessee Eastman data, reactor-cooling fault IDV4 is aligned, with 75% of attribution on its expected loop variables compared with 6% under the null. Condenser-cooling fault IDV5 receives more expected attribution than the null (0.22 versus 0.06) but is flagged for a small outside-set excess on the E-feed valve (0.026). Temporal features improve IDV5 recall and relevant attribution, yet outside-set flags remain. These may reflect downstream controller responses rather than shortcuts, which the available mechanism map cannot distinguish. To check both nulls under temporal dependence, we repeat the audit on fault-free data, placing a placebo onset within the first 30 hours of 100 randomly drawn runs per fault (three repetitions with different onsets and runs). Neither null produces a false alignment in the six placebo tests. One placebo IDV5 test is flagged under both nulls ($p = 0.02$ and 0.03), but its excess is 0.004, an order of magnitude below the reported IDV5 excess. Three repetitions can reveal gross miscalibration but cannot establish a false-flag rate.

When physical responses extend beyond the predefined feature set, unexpected attribution may have more than one plausible explanation.

b) Scalability and sample size: Table VII summarizes experiments with up to 3×10^5 observations. Without planted shortcuts, HDF, PWF, and OSF remain aligned across the tested sizes. The strong shortcut ($\sigma = 0.25$) is flagged in every evaluated run, including at the largest size, where the relevant-share test alone would pass despite substantial shortcut attribution. However, some larger datasets also produce small outside-set TWF flags, while weaker shortcuts can remain undetected. To calibrate an effect-size threshold, we audited 19 further generated datasets without shortcuts, using seeds disjoint from Table VII (10 at 10^4, 6 at 3×10^4, 3 at 10^5; $K = 200$). C_m falsely flags 0, 0, and 1 of these runs at the three sizes, and $C_m^{(3)}$ falsely flags 1, 2, and 2. Every false flag concerns TWF, so the false-flag rate grows with n as small real excesses become significant. The largest excess in any of these runs is 0.040 for C_m and 0.109 for $C_m^{(3)}$, which we adopt as thresholds. In 12 independent runs with planted sensors at 10^4, every flag exceeds these thresholds ($C_m \geq 0.18$, $C_m^{(3)} \geq 0.32$). C_m detects 2 of 3 runs at $\sigma = 0.25$ with one or three sensors. $C_m^{(3)}$ detects 2 of 3 with one sensor, 3 of 3 with three sensors, and 1 of 3 with three sensors at $\sigma = 0.35$. Applied to the runs in Table VII, the C_m threshold removes five of seven small TWF flags. The other two (0.047, both from one dataset at 3×10^4) exceed it, so the threshold reduces but does not eliminate false flags.

A five-fold audit requires $5(K + 1)$ model fits; one cross-fitted run took 115 seconds at 10^6 observations on one core. The largest dataset has only one seed. Statistical significance must therefore be read alongside effect size, cost, and the completeness of the mechanism description.

VI. CONCLUSION

This study presented a statistical audit to check whether predictive maintenance explanations agree with known failure mechanisms. The results underline why prediction accuracy alone is not enough to judge this agreement. On AI4I, explanations aligned with heat dissipation, power, and overstrain failure rules under the main settings, while the tool-wear result depended on how explanations were computed and which reference data were used. The choice of statistical comparison also mattered: a random-subset baseline suggested alignment even for a model trained on shuffled labels. In controlled experiments, accurate models could rely on fabricated sensors. The audit detected stronger shortcuts, and a grouped statistic extended detection to shortcuts spread over a few features at the cost of more false flags, but weaker shortcuts were missed. Tests on CWRU bearing recordings and Tennessee Eastman simulations supported alignment for inner- and outer-race bearing faults and the reactor-cooling fault under the main settings. Verdicts should be reported with their sensitivity to the model, explainer, attribution reference, and expected features. Agreement does not establish physical causality, and a model can still use shortcuts that the audit does not detect.

VII. LIMITATIONS AND FUTURE WORK

The proposed audit evaluates whether explanations agree with a predefined set of mechanism features under a specified attribution method and statistical reference. Significant outside-set attribution marks features for investigation but does not establish shortcut reliance or physical causality. Our evaluation uses simulated processes and laboratory recordings, so it does not establish performance on operating industrial equipment. Future work will validate the framework on maintenance data with verified failure causes and develop methods for defining expected feature sets in complex systems.

ACKNOWLEDGMENTS

This work used Bridges-2 at Pittsburgh Supercomputing Center through allocation CIS261821 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support

(ACCESS) program [38], which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

The authors used Claude (Anthropic), through Claude Code, for debugging and language editing. The research ideas, methods, experiments, and conclusions are the authors' own; they verified all AI-assisted material and take full responsibility for the content.

REFERENCES

[1] L. Cummins, A. Sommers, S. Bakhtiari Ramezani, S. Mittal, J. Jabour, M. Seale, and S. Rahimi, "Explainable predictive maintenance: A survey of current methods, challenges and opportunities," *IEEE Access*, vol. 12, pp. 57 574–57 602, 2024.

[2] Z. Allah Bukhsh, A. Saeed, I. Stipanovic, and A. G. Doree, "Predictive maintenance using tree-based classification techniques: A case of railway switches," *Transportation Research Part C: Emerging Technologies*, vol. 101, pp. 35–54, 2019.

[3] A. Jalali, A. Schindler, B. Haslhofer, and A. Rauber, "Machine learning interpretability techniques for outage prediction: A comparative study," *PHM Society European Conference*, vol. 5, no. 1, 2020.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[5] B. Steurtewagen and D. Van den Poel, "Adding interpretability to predictive maintenance by machine learning on sensor data," *Computers & Chemical Engineering*, vol. 152, p. 107381, 2021.

[6] S. Gawde, S. Patil, S. Kumar, P. Kamat, K. Kotecha, and S. Alfarhood, "Explainable predictive maintenance of rotating machines using LIME, SHAP, PDP, ICE," *IEEE Access*, vol. 12, pp. 29 345–29 361, 2024.

[7] M. L. Ndao, G. Youness, N. Niang, and G. Saporta, "Improving predictive maintenance: Evaluating the impact of preprocessing and model complexity on the effectiveness of eXplainable Artificial Intelligence methods," *Engineering Applications of Artificial Intelligence*, vol. 144, p. 110144, 2025.

[8] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[9] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, p. 1096, 2019.

[10] D. Janzing, L. Minorics, and P. Blöbaum, "Feature relevance quantification in explainable AI: A causal problem," in *Proc. 23rd Int. Conf. Artificial Intelligence and Statistics*, vol. 108, 2020, pp. 2907–2916.

[11] Y. Zhou, S. Booth, M. T. Ribeiro, and J. Shah, "Do feature attribution methods correctly attribute features?" in *Proc. AAAI Conf. Artificial Intelligence*, vol. 36, no. 9, 2022, pp. 9623–9633.

[12] A. Altmann, L. Tološi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010.

[13] "AI4I 2020 Predictive Maintenance Dataset," UCI Machine Learning Repository, 2020, DOI: https://doi.org/10.24432/C5HS5C.

[14] S. Matzka, "Explainable artificial intelligence for predictive maintenance applications," in *2020 Third Int. Conf. Artificial Intelligence for Industries (AI4I)*, 2020, pp. 69–74.

[15] W. Zhao, J. Ding, X. Huang, and Y. Zhang, "Research on milling machine predictive maintenance based on machine learning and SHAP analysis in intelligent manufacturing environment," arXiv:2512.01205, 2025, accepted at ICEIEC 2025.

[16] O. Obeten and A. Jimoh, "Physics-informed, cost-aware fault classification with comparative explainability for predictive maintenance: a SHAP–LIME agreement analysis," *Scientific Reports*, 2026, doi: 10.1038/s41598-026-64374-2.

[17] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[18] Case Western Reserve University, "Bearing data center: 12k drive end bearing fault data," https://engineering.case.edu/bearingdatacenter.

[19] C. C. Reinartz, M. Kulahci, and O. Ravn, "An extended Tennessee Eastman simulation dataset for fault-detection and decision support systems," *Computers & Chemical Engineering*, vol. 149, p. 107281, 2021.

[20] Q. Chen, X. Dong, K. Hu, K. Chen, Z. Peng, and G. Meng, "CS-SHAP: Extending SHAP to cyclic-spectral domain for better interpretability of intelligent fault diagnosis," arXiv:2502.06424, 2025.

[21] G. Gravanis, D. Kyriakou, S. Voutetakis, S. Papadopoulou, and K. Diamantaras, "Explainability for fault detection system in chemical processes," *arXiv:2602.16341*, 2026.

[22] I. Covert, S. Lundberg, and S.-I. Lee, "Explaining by removing: A unified framework for model explanation," *Journal of Machine Learning Research*, vol. 22, no. 209, pp. 1–90, 2021.

[23] H. Chen, J. D. Janizek, S. Lundberg, and S.-I. Lee, "True to the model or true to the data?" *arXiv preprint arXiv:2006.16234*, 2020.

[24] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values," *Artificial Intelligence*, vol. 298, p. 103502, 2021.

[25] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Machine Learning*, vol. 70, 2017, pp. 3319–3328.

[26] P. Sturmfels, S. Lundberg, and S.-I. Lee, "Visualizing the impact of feature attribution baselines," *Distill*, 2020.

[27] Y. Liu, S. Khandagale, C. White, and W. Neiswanger, "Synthetic benchmarks for scientific research in explainable machine learning," in *Proc. NeurIPS Datasets and Benchmarks Track*, 2021.

[28] C. Agarwal, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, and H. Lakkaraju, "OpenXAI: Towards a transparent evaluation of model explanations," in *Adv. Neural Inf. Process. Syst. (Datasets and Benchmarks Track)*, vol. 35, 2022.

[29] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," in *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 9737–9748.

[30] S. Da Veiga, "Global sensitivity analysis with dependence measures," *Journal of Statistical Computation and Simulation*, vol. 85, no. 7, pp. 1283–1305, 2015.

[31] M. Ojala and G. C. Garriga, "Permutation tests for studying classifier performance," *Journal of Machine Learning Research*, vol. 11, no. 62, pp. 1833–1863, 2010.

[32] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[33] D. Kim, H. Choi, M. J. A. Rasool, and G. Oh, "ϕ-test: Global feature selection and inference for Shapley additive explanations," arXiv:2512.07578, 2025.

[34] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[35] K. Debanath, P. Uprety, and P. Li, "Computational snapshot for auditing SHAP explanations against known failure mechanisms," Supplement paper/supplement.pdf and archive paper/artifacts/tuai-computation-5b97414ab7ef.zip at https://github.com/kowshik24/shap-mechanism-audit, 2026, snapshot ID: `5b97414ab7ef`; full SHA-256 and file hashes accompany the archive.

[36] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Int. Conf. Intelligent Computing*, 2005, pp. 878–887.

[37] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[38] T. J. Boerner, S. Deems, T. R. Furlani, S. L. Knuth, and J. Towns, "ACCESS: Advancing innovation: NSF's advanced cyberinfrastructure coordination ecosystem: Services & support," in *Practice and Experience in Advanced Research Computing (PEARC '23)*. ACM, 2023.