RiESTech
Recent in Engineering Science and Technology

MBI

Article

Auditing Physical Consistency in Engine Fault Detection Using Domain-Informed Monotonic XGBoost and Explainable Boosting Machine: A Case Study on a Public Marine Diesel Engine Dataset

Abdul Azis Abdillah [1,*]

[1] Department of Mechanical Engineering, School of Engineering, University of Birmingham, United Kingdom
* Correspondence: axa2072@student.bham.ac.uk

Abstract: Machine-learning models for engine fault detection are usually judged by accuracy, although a model can score well while learning relationships that contradict engine physics. This study proposes a physical-consistency audit that combines domain-informed monotonic constraints in XGBoost with the glass-box Explainable Boosting Machine (EBM). Three XGBoost configurations (unconstrained, domain-constrained and reverse-constrained) are compared, and their difference is summarised by a physical-consistency gap (PCG) with a confidence interval and a decision rule. The audit is applied to a public marine diesel engine dataset (IEEE DataPort; 19,533 records, six sensors). Its label is documented by the dataset creator (1 = no failure, 0 = failure), and the records are a synthetic extension of 174 shipboard observations. Under a leakage-safe protocol, the best of six classifiers reached a modest test ROC-AUC of 0.708. Engineering-consistent constraints had no measurable cost (test ΔAUC +0.0001), whereas reversed constraints reduced ROC-AUC by 0.015, giving PCG = −0.0154 (95% CI −0.0208 to −0.0101). SHAP and EBM agreed that fault risk falls with oil pressure and rises with oil temperature. Semi-synthetic labels, 30 null replicates and a positive control on the AI4I 2020 dataset showed that the audit separates consistent, inverted and non-physical labels; no null replicate was flagged. Coding the label the other way round, as in the first version of this study, produced the mirror-image gap and a flag, so the audit also detects a wrong label convention. Mixed tail evidence and the synthetic origin limit what the labels reveal about real engines.

Keywords: Engine fault detection; Marine diesel engine; Physical consistency; Monotonic constraints; Explainable boosting machine; SHAP; Label convention; Positive control

Citation: Abdillah, A. A. (2026). Auditing Physical Consistency in Engine Fault Detection Using Domain-Informed Monotonic XGBoost and Explainable Boosting Machine: A Case Study on a Public Marine Diesel Engine Dataset. Recent in Engineering Science and Technology, 4(4), 224–244. Retrieved from https://www.mbi-jounals.com/index.php/riestech/article/view/180.

Academic Editor: Yudhi Ariadi

Received: 1 October 2026
Accepted: 18 October 2026
Published: 31 October 2026

Publisher's Note: MBI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2026 by the authors. Licensee MBI, Jakarta, Indonesia. This article is an open access article distributed under MBI license (https://mbi-journals.com/licenses/by/4.0/).

1. Introduction

Condition-based and predictive maintenance aim to detect deteriorating equipment before it fails, reducing unplanned downtime and maintenance cost [1,2]. On ships, the main engine is the most critical item of machinery. Routinely logged variables such as engine speed, lubrication-oil pressure and temperature, fuel pressure and coolant pressure and temperature are natural inputs for data-driven fault detection [3]. Machine-learning (ML) classifiers have become the dominant approach for mapping these signals to a health state [4–6].

Most studies judge such models by predictive metrics alone, typically accuracy, F1-score or the area under the receiver operating characteristic curve (ROC-AUC). High accuracy, however, does not guarantee that a model has learned the right relationships. In a well-known medical example, a pneumonia-risk model learned that asthma lowers mortality risk [7]. This was an artefact of how asthmatic patients were treated, not a physiological effect. Models readily exploit such shortcuts [8], and image classifiers have been shown to rely on spurious "Clever Hans" cues [9]. Ross et al. formalised the aim of

models that are "right for the right reasons" [10]. In maintenance, technicians must trust and act on alerts. A detector that treats rising oil temperature as a sign of health would therefore be unacceptable, even with a competitive ROC-AUC [11].

Explainable AI (XAI) is often proposed as the remedy. SHapley Additive exPlanations (SHAP) [12,13] are widely used in predictive-maintenance studies [6]. In many applications, however, explanations are shown as importance rankings or dependence plots without a formal test of whether each learned direction agrees with engineering expectations. Theory-guided data science lists physical consistency as a requirement for scientific ML models [14]. Physics-informed ML embeds scientific knowledge into models [15], and monotonic constraints encode directional knowledge in tree ensembles [16]. Recent work has begun to use such constraints diagnostically. Koklev quantified the "price of monotonicity" as the performance change between unconstrained and constrained gradient boosting [17]. Abbas et al. refitted an XGBoost classifier with pre-specified monotonic constraints as a direction test [18], and Hsiao et al. corrected non-physical EBM shape functions with domain knowledge [19]. These studies price or repair a model. They do not provide a calibrated statistic that indicates whether the labels of a dataset agree with the assumed physics. This is the first gap addressed here.

The second gap concerns the data themselves. Datasheets for datasets were proposed because public data are often released without documentation of how they were collected and labelled [20]. Label noise and systematic labelling errors degrade classifiers and mislead their interpretation [21], and confident learning estimates label errors from model disagreement [22]. These methods do not exploit directional physical knowledge. A wrong label convention, in which the two classes are swapped, leaves accuracy and model confidence unchanged while reversing every learned direction, so it cannot be detected in this way. Data leakage can further inflate reported performance [23,24]. The public ship main-engine dataset studied here illustrates the problem [25]. Its repository record does not state which label value denotes a fault, although the creator's articles do [3,26]. Studies of this file report accuracies or areas under the curve between about 0.74 and above 0.98 [26–28], and the label convention is not always stated.

This study addresses both gaps with a physical-consistency audit built on two complementary models. The first is a three-way comparison of XGBoost [29] configurations that differ only in their monotonic constraints: unconstrained, constrained in the direction dictated by engineering knowledge, and constrained in the reversed direction. The second is the Explainable Boosting Machine (EBM) [30,31], a glass-box generalised additive model whose shape functions show the learned directions without any constraint. Relative to the work cited above, the distinctive elements are threefold. The first is the reversed-constraint control. The second is the physical-consistency gap (PCG), a statistic with a confidence interval and explicit decision rules. The third is a semi-synthetic calibration that shows which PCG values arise under consistent, inverted, partially reversed and non-physical labels. The audit is also checked on a positive-control dataset with documented failure mechanisms [32]. All experiments follow a leakage-safe protocol with repeated cross-validation, corrected statistical tests and a single evaluation on an isolated hold-out set.

2. Materials and Experiment Methods

2.1. Dataset, Label Convention and Data Origin

The study uses the public dataset "Predictive Maintenance on Ship's Main Engine using AI" [25], deposited on IEEE DataPort (DOI 10.21227/g3za-v415) and downloaded on 29 June 2026. The file engine_data.csv has 19,535 records with six continuous sensor variables (Table 1) and a binary label, Engine Condition. The repository record does not define the label values. The dataset creator defines them in two articles. Mohakul et al. state that "if engine condition = 0, the engine will fail and if engine condition = 1, no failure

will take place" [3]. Nurdin et al., in an article co-authored by the creator, define the label as healthy (1) and unhealthy (0) and state that the data describe an MTU Series 1400 marine diesel engine, with pressures in bar and temperatures in °C [26]. The 20 example records printed in that article match our file exactly, including their labels.

The models in this study therefore predict the fault class, defined as Fault = 1 − Engine Condition. Faults make up 36.9% of the cleaned records. The first version of this study used Engine Condition = 1 as the fault class. That coding is the reverse of the documented one, and Section 3.9 shows that the audit detects it.

The records are synthetic. Mohakul et al. describe how the data were produced [3]. Readings of the six sensors were transcribed from the logs of a ship's main engine, giving 174 records. Only three of them had Engine Condition = 0. These three were randomly oversampled to 51, and a Gaussian copula fitted with the Synthetic Data Vault [33] then generated 200,000 records. A companion conference paper also refers to "simulated data" [34]. The public file is smaller (19,535 records, 36.9% with Engine Condition = 0, against 34.3% in the 200,000 records), and how it was derived from the generated data is not documented. Its inter-sensor correlations are close to zero (Table 2). All pairwise Pearson correlations are below 0.074 in magnitude, and the correlation between oil pressure and engine speed is +0.025. In a running engine, oil pressure depends on speed and oil temperature, so the sensors were probably generated almost independently of one another. Two consequences follow. First, the audit tests the data-generating process and its labelling rather than the physics of a real engine. Second, the fault class rests on three genuine fault observations.

2.2. Preprocessing and Cleaning

The data contained no missing values and no duplicate records. One fixed domain rule was applied before data partitioning: records with a coolant temperature above 110 °C were removed. Two records were affected, with 118.4 °C (labelled fault) and 195.5 °C (labelled no fault). The next-highest value is 95.9 °C. A reading of 195.5 °C is far above the boiling range of a pressurised cooling circuit and was treated as a sensor error. A reading of 118.4 °C could reflect genuine overheating; it was removed only so that one consistent rule is applied, and Appendix Table A2 shows the audit with either record kept. Because the rule is not estimated from data, applying it before the split cannot leak information.

Minimum pressures are close to zero (Table 1), and 118 records report an oil pressure below 0.5 bar, which is incompatible with a running engine. Their other sensor readings are typical of the whole file, and their fault rate (0.356) is close to the base rate. This pattern suggests generated values outside the physical range rather than lubrication failures. Excluding these records did not change the audit result (Table A2). No feature scaling was applied to tree-based models; standardisation for logistic regression was embedded in a pipeline and fitted on training folds only. No resampling was used, because the class ratio (63:37) is only moderately imbalanced and resampling would alter the conditional distributions that the audit examines.

Table 1. Sensor variables after cleaning (n = 19,533), their association with the fault class and the expected direction of fault risk used for the monotonic constraints.

Variable	Mean	SD	Min	Max	Mean (healthy)	Mean (fault)	r_{pb} (t)	Expected fault risk
Engine speed (rpm)	791.26	267.61	61.000	2239.00	736.32	885.01	+0.268 (+38.9)	unconstrained
Lub. oil pressure (bar)	3.30	1.02	0.003	7.27	3.35	3.22	−0.061 (−8.6)	↑ as pressure ↓ (−1)
Fuel pressure (bar)	6.66	2.76	0.003	21.14	6.90	6.24	−0.116 (−16.3)	unconstrained
Coolant pressure (bar)	2.34	1.04	0.003	7.48	2.32	2.37	+0.024 (+3.4)	unconstrained
Lub. oil temperature (°C)	77.64	3.11	71.322	89.58	77.42	78.02	+0.094 (+13.1)	↑ as temp. ↑ (+1)
Coolant temperature (°C)	78.42	6.14	61.673	95.86	78.20	78.80	+0.047 (+6.6)	↑ as temp. ↑ (+1)

Units follow Nurdin et al. [26]. Healthy: Engine Condition = 1; fault: Engine Condition = 0. r_{pb}: point-biserial correlation with the fault class; t: corresponding t statistic (df = 19,531); $|t| > 3.3$ corresponds to $p < 0.001$.

Table 2. Pearson correlation matrix of the six sensors and the fault class (cleaned data).

	Speed	Oil p.	Fuel p.	Cool. p.	Oil T	Cool. T	Fault
Speed	1						
Oil p.	0.025	1					
Fuel p.	−0.002	0.044	1				
Cool. p.	−0.025	−0.009	0.033	1			
Oil T	0.052	−0.008	−0.025	−0.021	1		
Cool. T	0.031	−0.061	−0.043	0.035	0.074	1	
Fault	0.268	−0.061	−0.116	0.024	0.094	0.047	1

Largest absolute inter-sensor correlation: 0.074. Spearman correlations (supplementary results file) give the same picture.

2.3. Leakage-Safe Experimental Protocol

Figure 1 summarises the protocol. After cleaning, the data were split once into a stratified training set (80%, 15,626 records) and a hold-out test set (20%, 3,907 records). Both retain the 36.9% fault prevalence. The test set was isolated immediately and scored once, at the end. All operations that learn from data were carried out on the training set only [23,35]: scaling, hyper-parameter tuning, model comparison and threshold selection. Hyper-parameters were tuned by randomised search with three-fold stratified cross-validation (CV) on the training set, maximising ROC-AUC (Table A1). Models were then compared with 5 × 3 repeated stratified CV on the same training set.

This benchmark design is not nested: the tuning data are reused for the comparison, which can bias CV estimates optimistically [35]. The hold-out results are therefore treated as the primary performance evidence. For the three XGBoost configurations that form the audit, a fully nested CV was also run, in which each configuration was re-tuned inside every outer training fold. For threshold-dependent metrics, the threshold of each model was chosen by maximising Youden's J [36] on five-fold out-of-fold training predictions and then frozen before the test set was scored.

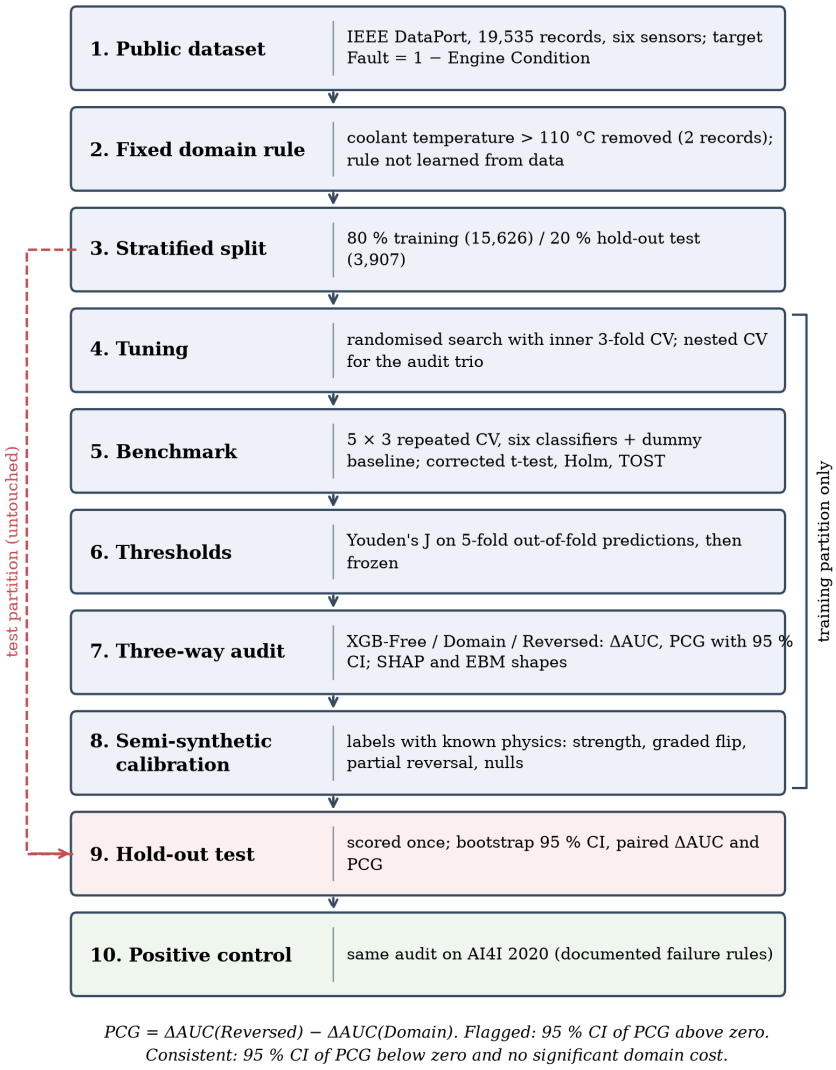


Figure 1. Leakage-safe experimental protocol. The hold-out test set is separated at step 3 and scored once (step 9). Tuning, benchmarking, threshold selection, the audit and the semi-synthetic calibration use the training partition only; the positive control repeats the audit on the AI4I 2020 dataset.

2.4. Benchmark Classifiers

Six classifiers were benchmarked with scikit-learn [37] and the respective libraries: logistic regression (LR) with L2 regularisation; random forest (RF) [38]; LightGBM [39]; CatBoost [40]; unconstrained XGBoost (XGB-Free) [29]; and EBM [30,31]. A majority-class dummy classifier, which ignores the sensors and always predicts the most frequent class (healthy), served as the minimum reference that any useful model must exceed; it is not part of any other model. The two constrained XGBoost configurations of the audit (Section 2.5) were evaluated alongside the benchmark. The tuned hyper-parameters are listed in Table A1. The two models central to the audit are described below.

2.5. Domain-Informed Monotonic XGBoost

XGBoost builds an additive ensemble of regression trees by gradient boosting. A monotonic constraint $d_f \in \{-1, 0, +1\}$ on feature f restricts every split on that feature. The ensemble output $F(\mathbf{x})$, the log-odds of a fault, is then non-decreasing ($d_f = +1$) or non-increasing ($d_f = -1$) in x_f when all other features $\mathbf{x}_{-f}$ are held fixed:

$$d_f = +1: \quad x_f \le x'_f \;\Rightarrow\; F(x_f, \mathbf{x}_{-f}) \le F(x'_f, \mathbf{x}_{-f}) \tag{1a}$$

$$d_f = -1: \quad x_f \le x'_f \;\Rightarrow\; F(x_f, \mathbf{x}_{-f}) \ge F(x'_f, \mathbf{x}_{-f}) \tag{1b}$$

The domain directions encode the dominant failure mechanisms of the lubrication and cooling systems of a diesel engine [41]. Low lubrication-oil pressure indicates insufficient oil flow to the bearings ($d = -1$), and high oil and coolant temperatures indicate overheating ($d = +1$). These directions are not universal. A low coolant temperature can indicate a thermostat stuck open, and a high oil pressure can indicate a stuck relief valve or cold, viscous oil. Engine speed, fuel pressure and coolant pressure were left unconstrained because both abnormally low and abnormally high values can indicate faults. The same argument applies, more weakly, to the three constrained sensors. In addition, the observed oil-temperature range (71–90 °C) and most of the coolant-temperature range (62–96 °C) lie inside normal operating windows. Manufacturer alarm set points for this engine are not publicly available and were therefore not used.

Monotonicity is therefore treated as a hypothesis about the dominant mechanism, not as a law. Three analyses test it: constraints applied to one sensor at a time, a tail analysis restricted to physically unambiguous extremes (Section 2.7), and EBM shape functions that are free to be non-monotone. Three configurations were defined: XGB-Free (no constraints), XGB-Domain (the directions above) and XGB-Reversed (all directions negated). In the main audit all three reuse the hyper-parameters tuned for XGB-Free, so that differences are attributable to the constraints alone. Because shared hyper-parameters may favour XGB-Free, the constrained configurations were also tuned separately with the same search, and the audit was repeated under nested CV.

2.6. Explainable Boosting Machine

EBM is a generalised additive model with pairwise interactions (GA²M) fitted by cyclic gradient boosting [30,31]. Each boosting round fits a shallow tree to a single feature in round-robin fashion with a small learning rate. The effects of different features are therefore learned separately and summed:

$$g(\mathrm{E}[y]) = \beta_0 + \Sigma_f\, s_f(x_f) + \Sigma_{(i,j)}\, s_{ij}(x_i, x_j) \tag{2}$$

Here g is the logit link, s_f is the shape function of feature f, and s_{ij} are pairwise interaction terms. Because each s_f can be plotted directly, the direction of every learned main effect can be read without post-hoc explanation. EBM serves as a benchmark model and as an unconstrained view of the learned directions. The tuned EBM included 10 interaction terms. With six features there are only 15 possible pairs, so interactions could absorb part of the main effects; an EBM without interaction terms was therefore also fitted.

2.7. Physical-Consistency Audit

The audit has three quantitative components. The first is the cost of each constraint configuration c relative to the unconstrained model, and the resulting physical-consistency gap (PCG):

$$\Delta\mathrm{AUC}_c = \mathrm{AUC}_c - \mathrm{AUC}_{\mathrm{Free}}, \quad c \in \{\mathrm{Domain}, \mathrm{Reversed}\} \tag{3}$$

$$\mathrm{PCG} = \Delta\mathrm{AUC}_{\mathrm{Reversed}} - \Delta\mathrm{AUC}_{\mathrm{Domain}} = \mathrm{AUC}_{\mathrm{Reversed}} - \mathrm{AUC}_{\mathrm{Domain}} \tag{4}$$

If the labels follow the assumed physics, domain constraints should cost little ($\Delta\mathrm{AUC}_{\mathrm{Domain}} \approx 0$) and reversed constraints should be costly, giving $\mathrm{PCG} < 0$. A positive PCG indicates that the labels are better described by the reversed physics. In CV, ΔAUC and PCG were computed in each of the $k = 15$ folds. Their 95% confidence intervals use the corrected resampled variance of Nadeau and Bengio [42]:

$$\overline{\mathrm{PCG}} \pm t_{0.975,\,k-1} \cdot [(1/k + n_2/n_1)\, s^2]^{1/2} \tag{5}$$

where n_1 and n_2 are the training and validation sizes of a fold and s^2 is the sample variance of the fold-wise PCG values. On the test set, 2,000 bootstrap resamples of the test records were drawn once and shared by all models (paired bootstrap). ΔAUC and PCG

were recomputed on every resample, and the 2.5th and 97.5th percentiles give the 95% CI [43]. **Decision rules** use the CV estimates. A dataset is *flagged* as inconsistent with the assumed physics when the lower bound of the 95% CI of PCG lies above zero. It is read as *consistent* when the upper bound lies below zero and, in addition, the domain constraints cost no significant discrimination (upper bound of the 95% CI of $\Delta AUC_{Domain} \geq 0$). The second condition is needed because a non-physical rule can make both constraint sets costly and so also produce a negative PCG. All other results are *unresolved*. The error rates of both rules were estimated from null scenarios (Section 2.8).

The second component is directional agreement. For each feature, the learned direction was measured by the Spearman correlation ρ_f between the feature value and its contribution: the SHAP value of XGB-Free computed with TreeSHAP [13], or the EBM term contribution. Because ρ captures only the sign of a monotone trend, an effect size was added:

$$\Delta_f = \text{mean}\{\varphi_f(\mathbf{x}) : x_f \geq P_{95}\} - \text{mean}\{\varphi_f(\mathbf{x}) : x_f \leq P_5\} \quad (6)$$

where φ_f is the contribution in log-odds of a fault and P_5 and P_{95} are the 5th and 95th percentiles of the feature on the test set. Confidence intervals for ρ and Δ come from 1,000 bootstrap resamples of the test records. A feature is consistent when $\text{sign}(\rho_f) = d_f$ for SHAP, EBM and EBM without interactions, and inconsistent when all three signs oppose d_f. Changes in each feature's share of mean absolute SHAP value were also recorded.

The third component is model-free. A standardised logistic regression with Wald 95% CIs was fitted to all cleaned records as a simple baseline. In addition, the fault rate was computed in physically unambiguous extremes: oil pressure below 1.0 and 0.5 bar or above 6.0 bar, oil temperature above 85 °C, coolant temperature above 93 °C or below 65 °C, fuel pressure below 2.0 bar and engine speed below 300 rpm. These limits are conservative extremes of the observed ranges, not manufacturer alarm set points. If the labels are physical, each of these high-risk regions should show a fault rate above the base rate. Wilson 95% intervals [44] and exact binomial tests against the base rate were used.

2.8. Semi-Synthetic Calibration of the Audit

To establish which audit results to expect under known conditions, semi-synthetic labels were generated on the training partition only, keeping the real sensor values. With features standardised by the training mean and standard deviation ($\mathbf{z}$), fault labels were drawn as Bernoulli variables with probability

$$P(y = 1 \mid \mathbf{z}) = \sigma(\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{z}), \quad \sigma(u) = 1 / (1 + e^{-u}) \quad (7)$$

The coefficient vector was derived from the data rather than chosen freely. For the three unconstrained sensors, $\boldsymbol{\beta}$ equals the standardised logistic-regression coefficients of the real training labels: +0.583 (engine speed), −0.261 (fuel pressure) and +0.066 (coolant pressure). For the three constrained sensors, $\beta_f = k \cdot d_f \cdot |\hat{\beta}_f|$ keeps the magnitude of the real coefficient ($|\hat{\beta}|$ = 0.145, 0.170 and 0.046 for oil pressure, oil temperature and coolant temperature) but takes the sign of the domain direction, scaled by a strength factor k. The intercept β_0 was solved numerically so that the expected fault rate equals the real rate (36.9%); for $k = 1$, $\beta_0 = -0.581$.

Five groups of scenarios were examined: (a) consistent labels (k = 1) and their inversion ($1 - y$); (b) signal strengths $k = 0.5$, 2 and 4, consistent and inverted, which bracket the real effect size; (c) a graded label flip, in which a random fraction q = 0.1, …, 0.9 of consistent labels was inverted, for $k = 1$ and $k = 2$; (d) partial reversal, in which the physical sign was reversed for one or two of the three constrained sensors; and (e) three null scenarios with ten replicates each. In null N1, labels depend only on the unconstrained sensors. In N2, a non-physical U-shaped rule makes risk rise at both ends of every constrained sensor. In N3, the real labels are randomly permuted. Every scenario used the same 5 × 3 repeated CV and the same XGBoost hyper-parameters as the real-data audit.

2.9. Positive Control on a Dataset with Documented Failure Mechanisms

The AI4I 2020 predictive-maintenance dataset [32] contains 10,000 records and documents the rules that generate its failures (machine-failure rate 3.39%). Heat-dissipation failure occurs when the difference between process and air temperature is below 8.6 K and the rotational speed is below 1380 rpm; tool-wear and overstrain failures increase with tool wear. Fault risk is therefore non-increasing in the temperature difference ($d = -1$) and non-decreasing in tool wear ($d = +1$). Product type, air temperature, speed and torque were left unconstrained, because power failure occurs at both low and high power. The same protocol was applied: an 80/20 stratified split, XGBoost tuning on the training part, the three-way audit with 5 × 3 CV, and a paired bootstrap on the test part. The audit was run with the documented labels and with inverted labels. Unlike the ship dataset, where the truth is unknown, this dataset shows whether the audit returns the correct answer when the truth is known.

2.10. Evaluation Metrics and Statistical Testing

Threshold-independent metrics were ROC-AUC, the area under the precision-recall curve (PR-AUC) [45] computed with the fault class and, separately, the healthy class as the positive class, and the Brier score. The Brier score is compared with a prior-probability baseline that predicts the training prevalence for every record, and the Brier skill score is $BSS = 1 - \text{Brier}/\text{Brier}_{\text{prior}}$. Calibration slope and calibration-in-the-large intercept were obtained by logistic recalibration of the predicted log-odds [46]. At the frozen Youden threshold, balanced accuracy, the Matthews correlation coefficient (MCC) [47], and precision, recall and F1-score for both classes were reported.

Test-set ROC-AUC confidence intervals are percentile intervals from 2,000 bootstrap resamples [43]. Differences in CV ROC-AUC were tested with the Nadeau–Bengio corrected resampled *t*-test [42], and comparisons against the best model were adjusted with the Holm procedure [48]. Because a non-significant difference does not show equivalence, equivalence was tested with two one-sided tests (TOST) [49] using the same corrected variance, and with the Bayesian correlated *t*-test of Benavoli et al. [50]. Both use a region of practical equivalence (ROPE) of ±0.01 ROC-AUC, fixed before the analysis as a difference too small to matter for maintenance decisions at the present level of discrimination. All analyses used Python 3.11 with fixed random seed 42; package versions are listed in the supplementary environment file. The code and results are archived on Zenodo (https://doi.org/10.5281/zenodo.23728914).

3. Results and Discussion

3.1. Data Characteristics and Simple Baselines

Table 1 and Figure 2 show that the two classes overlap heavily for every sensor. Engine speed has the strongest association with the fault class ($r_{pb} = +0.268$), followed by fuel pressure (−0.116) and lubrication-oil temperature (+0.094). The marginal associations of the three constrained sensors follow the expected fault physics: fault records have slightly lower oil pressure and slightly higher oil and coolant temperatures. With 19,533 records these differences are highly significant ($|t| = 8.6$, 13.1 and 6.6). Fault records also run at higher engine speed; speed was left unconstrained.

A standardised logistic regression gives the same picture (Table 3). The coefficients of oil pressure (−0.138), oil temperature (+0.167) and coolant temperature (+0.055) all have the expected sign, and their 95% CIs exclude zero. Under the reversed coding all these signs would be reversed, so a simple baseline is also sensitive to the label convention. The audit adds three things. First, it measures consistency in predictive terms, as the discrimination lost when the physics is enforced or reversed, and it does so with a flexible model that allows non-linear and conditional effects. Second, the semi-synthetic

calibration and the null scenarios show what size of PCG to expect under known label mechanisms, which a coefficient sign cannot. Third, the reversed-constraint control separates a consistent direction from a mere lack of signal: sensors without signal give PCG ≈ 0, not PCG < 0.

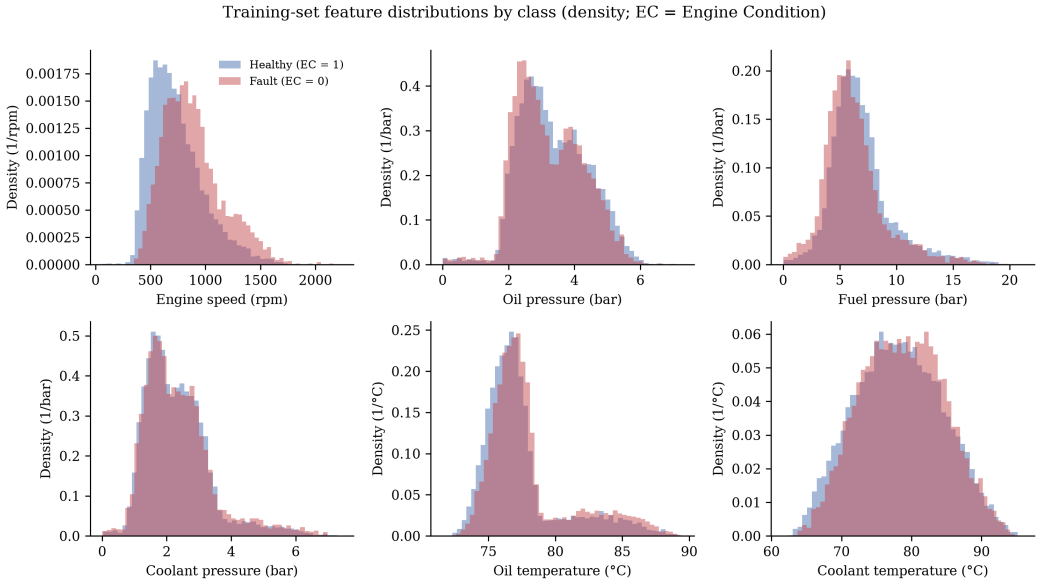


Figure 2. Training-set distributions (densities) of the six sensor variables by class. All variables overlap heavily between the two classes.

Table 3. Standardised logistic regression of the fault class on the six sensors (all cleaned records).

Variable	Coefficient per SD	95% CI	*p*	Expected sign	Verdict
Engine speed (rpm)	+0.584	[+0.552, +0.616]	<0.001	free	n/a
Lub. oil pressure (bar)	−0.138	[−0.169, −0.107]	<0.001	−	consistent
Fuel pressure (bar)	−0.268	[−0.301, −0.235]	<0.001	free	n/a
Coolant pressure (bar)	+0.078	[+0.047, +0.108]	<0.001	free	n/a
Lub. oil temperature (°C)	+0.167	[+0.137, +0.198]	<0.001	+	consistent
Coolant temperature (°C)	+0.055	[+0.024, +0.085]	<0.001	+	consistent

Wald confidence intervals; coefficients in log-odds of a fault per standard deviation of the sensor.

3.2. Tail Analysis in Physically Unambiguous Extremes

Table 4 and Figure 3 give the fault rate in extreme regions where engine physics is unambiguous (base rate 0.369). The evidence is mixed. The oil temperature > 85 °C and fuel pressure < 2.0 bar regions show a fault rate significantly above the base rate, as physics predicts. The coolant temperature < 65 °C and engine speed < 300 rpm regions show a fault rate significantly below it. The remaining regions (oil pressure < 1.0 bar, oil pressure < 0.5 bar, oil pressure > 6.0 bar and coolant temperature > 93 °C) are inconclusive. In particular, the low-oil-pressure tail, the most direct indicator of lubrication failure, is close to the base rate. These extremes contain few records and, given the synthetic origin of the data (Section 2.1), lie where the generator had the least information from the 174 source records. The tails therefore do not show that the labels are physically valid throughout the sensor range, only that the bulk evidence examined by the audit is not contradicted by them.

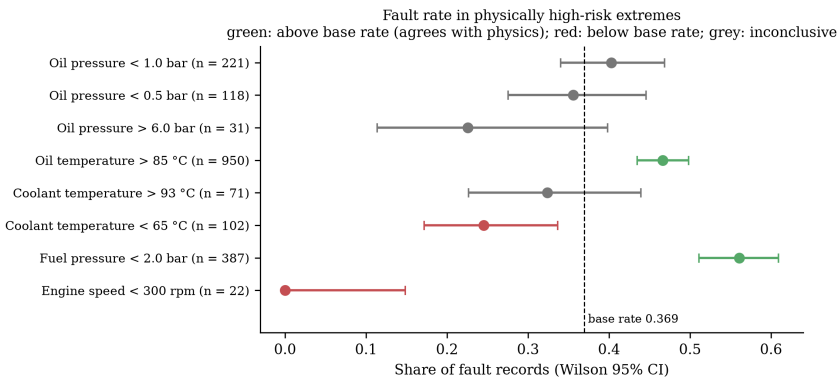


Figure 3. Fault rate in physically unambiguous extremes with Wilson 95% confidence intervals. The dashed line marks the base rate.

Table 4. Fault rate in physically unambiguous extremes (all cleaned records).

Region	*n*	Fault rate [95% CI]	*p* (vs base rate)	Physics
Oil pressure < 1.0 bar	221	0.403 [0.340, 0.469]	0.329	inconclusive
Oil pressure < 0.5 bar	118	0.356 [0.275, 0.446]	0.849	inconclusive
Oil pressure > 6.0 bar	31	0.226 [0.114, 0.398]	0.135	inconclusive
Oil temperature > 85 °C	950	0.466 [0.435, 0.498]	<0.001	agrees
Coolant temperature > 93 °C	71	0.324 [0.227, 0.439]	0.463	inconclusive
Coolant temperature < 65 °C	102	0.245 [0.172, 0.337]	0.010	contradicts
Fuel pressure < 2.0 bar	387	0.561 [0.511, 0.609]	<0.001	agrees
Engine speed < 300 rpm	22	0.000 [0.000, 0.149]	<0.001	contradicts

Base rate 0.369. Physics: Wilson interval entirely above (agrees) or below (contradicts) the base rate; otherwise inconclusive. Exact binomial test against the base rate.

3.3. Predictive Performance of the Benchmark

Table 5 and Figure 4a report the repeated-CV results on the training set. All learning models clearly outperformed the majority-class baseline (all $p < 0.001$), but the achievable discrimination is modest, with ROC-AUC around 0.70. EBM ranked first (0.7048 ± 0.0056). Non-significant differences do not establish equivalence, so the equivalence tests are decisive here. Within a margin of ±0.01, CatBoost, XGB-Free, LightGBM, XGB-Domain and Random Forest were statistically equivalent to EBM (Holm-adjusted TOST $p < 0.05$). Equivalence could not be shown for Logistic Regression and XGB-Reversed.

The hold-out results (Table 6, Figure 4b,c) agree with the CV ranking. CatBoost achieved the highest test ROC-AUC (0.708; 95% CI 0.691–0.725), and XGB-Reversed the lowest among the learning models (0.690). Test values are within about 0.01 of their CV counterparts, so there is no sign of overfitting or leakage. The glass-box EBM matches the black-box ensembles, in line with the argument that interpretable models need not sacrifice accuracy on tabular data [11].

A ROC-AUC near 0.70 limits practical use. At the chosen thresholds roughly a third of the records are misclassified (Table 7), which would be too many for alarm generation on a real engine. Published studies of the same data report much higher figures, for example 97% accuracy for a neural network trained on the 174 source records [26] and AUCs above 0.98 after oversampling [28]. Those figures are not directly comparable with ours, which use all records, the original labels and a single evaluation on an untouched hold-out set. The present study does not propose a deployable detector. Its aim is to test whether the relationships that any detector would learn from these labels are physically plausible, and this question can be answered at modest discrimination.

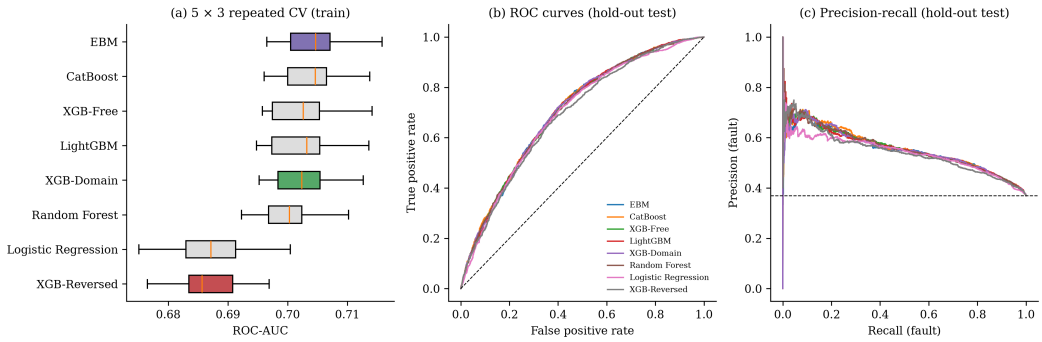


Figure 4. Predictive performance: (a) ROC-AUC over 15 repeated CV folds on the training set; (b) ROC curves and (c) precision-recall curves for the fault class on the hold-out test set. The dashed line in (c) marks the fault prevalence.

Table 5. Five-fold, three-times repeated cross-validation on the training set, ordered by mean ROC-AUC, with difference, equivalence and Bayesian tests against the best model.

Model	ROC-AUC (mean ± SD)	PR-AUC (fault)	PR-AUC (healthy)	Brier	ΔAUC vs. EBM [95% CI]	*p* (Holm)	*p* TOST (Holm)	P(ROPE)
EBM	0.7048 ± 0.0056	0.561	0.802	0.205	—	—	—	—
CatBoost	0.7040 ± 0.0054	0.561	0.800	0.206	−0.0008 [−0.0021, +0.0006]	0.241	<0.001	1.00
XGB-Free	0.7027 ± 0.0052	0.558	0.799	0.206	−0.0021 [−0.0040, −0.0002]	0.071	<0.001	1.00
LightGBM	0.7023 ± 0.0054	0.557	0.799	0.206	−0.0025 [−0.0044, −0.0007]	0.032	<0.001	1.00
XGB-Domain	0.7022 ± 0.0052	0.557	0.799	0.206	−0.0026 [−0.0042, −0.0010]	0.015	<0.001	1.00
Random Forest	0.7000 ± 0.0051	0.555	0.797	0.207	−0.0048 [−0.0073, −0.0023]	0.005	<0.001	1.00
Logistic Regression	0.6873 ± 0.0069	0.541	0.783	0.211	−0.0175 [−0.0208, −0.0143]	<0.001	1.000	0.00
XGB-Reversed	0.6868 ± 0.0062	0.537	0.787	0.210	−0.0180 [−0.0234, −0.0127]	<0.001	1.000	0.00
Dummy	0.5000 ± 0.0000	0.369	0.631	0.369	—	—	—	—

Corrected resampled *t*-test [42] with Holm adjustment; TOST and Bayesian correlated *t*-test with a ROPE of ±0.01 ROC-AUC; P(ROPE): posterior probability of practical equivalence.

Table 6. Hold-out test performance (n = 3,907): discrimination and calibration.

Model	ROC-AUC [95% CI]	PR-AUC (fault)	PR-AUC (healthy)	Brier	BSS	Cal. slope	Cal. intercept
CatBoost	0.708 [0.691–0.725]	0.557	0.802	0.205	+0.121	1.04	+0.012
EBM (no interactions)	0.707 [0.689–0.723]	0.553	0.801	0.205	+0.119	1.00	+0.011
LightGBM	0.706 [0.689–0.723]	0.555	0.803	0.205	+0.119	0.99	+0.011
EBM	0.706 [0.689–0.723]	0.552	0.801	0.205	+0.118	0.97	+0.011
XGB-Domain	0.706 [0.689–0.722]	0.554	0.801	0.205	+0.119	1.00	+0.011
XGB-Free	0.706 [0.689–0.722]	0.555	0.801	0.205	+0.119	1.02	+0.014
Random Forest	0.704 [0.686–0.721]	0.556	0.799	0.207	+0.113	1.25	+0.013
Logistic Regression	0.696 [0.679–0.713]	0.539	0.790	0.210	+0.098	1.08	+0.003
XGB-Reversed	0.690 [0.673–0.708]	0.542	0.790	0.209	+0.102	0.98	+0.018
Dummy	0.500	0.370	0.630	0.370	−0.586	—	—

BSS: Brier skill score relative to the prior-probability baseline (Brier = 0.233). Calibration slope and intercept from logistic recalibration (ideal values 1 and 0).

3.4. Threshold-Dependent Performance and Calibration

The dummy classifier predicts the majority class, healthy, for every record. It therefore never detects a fault (recall 0.000) yet reaches an accuracy equal to the healthy share, which is why accuracy is not reported as a headline metric. For the learning models, Table 7 reports precision, recall and F1-score for both classes. For EBM, the fault class has precision 0.509 and recall 0.713, and the healthy class precision 0.780 and recall 0.598. The Youden criterion, MCC and balanced accuracy do not depend on which class is called positive, which is why they were used for threshold selection. Criteria such as the recall-

weighted F2-score do depend on it: in the first version of this study, where the majority class was coded as the fault class, F2-optimal thresholds collapsed every model to the trivial all-majority prediction.

The dummy Brier score (0.370) reflects hard majority-class predictions. Against the fairer prior-probability baseline (Brier = 0.233), the Brier skill of the learning models is only 0.10–0.12. Calibration slopes range from 0.97 to 1.25 and intercepts are close to zero (Table 6), so the models are reasonably calibrated, but their probabilistic skill is limited (Figure 5).

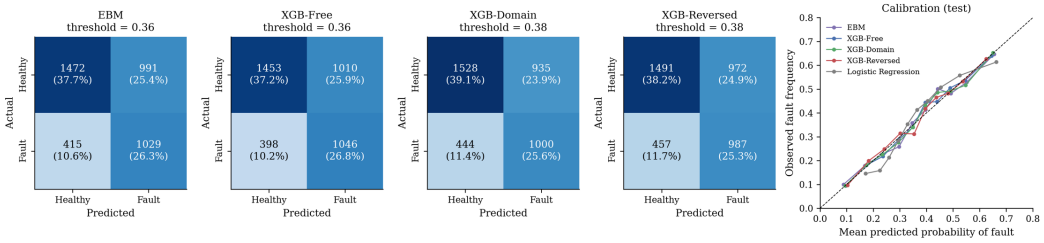


Figure 5. Confusion matrices on the hold-out test set at the frozen Youden thresholds for EBM and the three XGBoost configurations (common colour scale; fault = positive class), and calibration curves (quantile bins) for selected models.

Table 7. Hold-out test performance at the frozen Youden thresholds, for both classes.

				Fault class			Healthy class		
Model	**Thr.**	**Bal. acc.**	**MCC**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
CatBoost	0.37	0.659	0.307	0.514	0.712	0.597	0.782	0.606	0.683
EBM (no interactions)	0.38	0.658	0.304	0.520	0.688	0.592	0.774	0.627	0.693
LightGBM	0.36	0.659	0.308	0.509	0.732	0.601	0.789	0.586	0.673
EBM	0.36	0.655	0.300	0.509	0.713	0.594	0.780	0.598	0.677
XGB-Domain	0.38	0.656	0.302	0.517	0.693	0.592	0.775	0.620	0.689
XGB-Free	0.36	0.657	0.304	0.509	0.724	0.598	0.785	0.590	0.674
Random Forest	0.40	0.649	0.288	0.519	0.652	0.578	0.760	0.645	0.698
Logistic Regression	0.33	0.648	0.288	0.495	0.740	0.593	0.785	0.557	0.652
XGB-Reversed	0.38	0.644	0.279	0.504	0.683	0.580	0.765	0.605	0.676
Dummy	0.50	0.500	0.000	0.000	0.000	0.000	0.630	1.000	0.773

Thr.: threshold on the predicted fault probability, selected by Youden's J on out-of-fold training predictions.

3.5. Physical-Consistency Audit

The audit results are given in Table 8 and Figure 6. Imposing the engineering-consistent constraints changed ROC-AUC by only −0.0005 in CV (95% CI [−0.0023, +0.0013], p = 0.567) and by +0.0001 on the test set ([−0.0021, +0.0022]). Imposing the reversed constraints reduced it by 0.0159 in CV ([−0.0210, −0.0109], $p < 0.001$) and by 0.0153 on the test set ([−0.0233, −0.0073]). The physical-consistency gap is therefore negative: PCG = −0.0154 in CV ([−0.0208, −0.0101]) and −0.0154 on the test set ([−0.0232, −0.0073]). Under the decision rules of Section 2.7 the dataset is read as consistent with the assumed physics.

The result does not depend on shared hyper-parameters. With separately tuned constrained models, PCG = −0.0139 ([−0.0187, −0.0090]); under fully nested CV, PCG = −0.0133 ([−0.0175, −0.0091]). Constraining one sensor at a time (Table 9) gives single-sensor gaps of −0.0059 for oil pressure, −0.0077 for oil temperature, −0.0009 for coolant temperature. Their sum (−0.0145) is close to the joint PCG.

The consistency is carried by oil pressure and oil temperature, whereas coolant temperature alone gives an unresolved result; its learned direction is examined in Section 3.6.

The attribution analysis explains the mechanism (Figure 6b). Under reversed constraints, the shares of mean absolute SHAP value for oil pressure and oil temperature fell from 12.0% and 14.3% to 0.4% and 0.1%, and coolant temperature changed from 4.8% to 1.9%. The model compensated by relying more on engine speed and fuel pressure.

When forced to contradict the assumed physics, the model found no usable signal in the constrained sensors and discarded them. Under domain constraints the attribution profile was virtually identical to that of the unconstrained model, indicating that XGB-Free had already learned the physically expected directions.

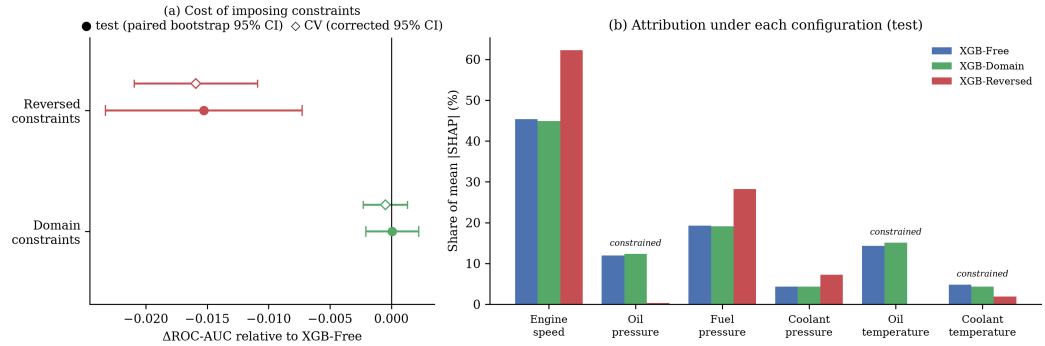


Figure 6. Physical-consistency audit. (a) Change in ROC-AUC caused by domain and reversed constraints relative to XGB-Free (filled markers: test-set paired bootstrap mean with 95% CI; open markers: CV mean with corrected 95% CI). (b) Share of mean absolute SHAP value per feature on the test set under each configuration.

Table 8. Physical-consistency audit: cost of constraints relative to XGB-Free and the physical-consistency gap (PCG, Equation (4)), under shared, separate and nested tuning.

Tuning	Quantity	CV estimate [95% CI]	CV p	Test estimate [95% CI]
Shared hyper-parameters	ΔAUC Domain	−0.0005 [−0.0023, +0.0013]	0.567	+0.0001 [−0.0021, +0.0022]
	ΔAUC Reversed	−0.0159 [−0.0210, −0.0109]	<0.001	−0.0153 [−0.0233, −0.0073]
	PCG	−0.0154 [−0.0208, −0.0101]	<0.001	−0.0154 [−0.0232, −0.0073]
Separately tuned	ΔAUC Domain	−0.0002 [−0.0024, +0.0019]	0.810	—
	ΔAUC Reversed	−0.0141 [−0.0190, −0.0093]	<0.001	—
	PCG	−0.0139 [−0.0187, −0.0090]	<0.001	−0.0125 [−0.0202, −0.0047]
Nested CV	ΔAUC Domain	−0.0004 [−0.0017, +0.0010]	0.562	—
	ΔAUC Reversed	−0.0137 [−0.0181, −0.0093]	<0.001	—
	PCG	−0.0133 [−0.0175, −0.0091]	<0.001	—

CV: 5 × 3 repeated CV on the training set with corrected 95% CI and corrected resampled t-test; test: paired bootstrap (2,000 resamples) on the hold-out set. Nested CV: each configuration re-tuned inside every outer fold.

Table 9. Audit with one sensor constrained at a time (5 × 3 CV, shared hyper-parameters).

Constrained sensor	ΔAUC Domain [95% CI]	ΔAUC Reversed [95% CI]	PCG [95% CI]	Reading
Oil pressure	−0.0001 [−0.0010, +0.0008]	−0.0061 [−0.0092, −0.0030]	−0.0059 [−0.0089, −0.0030]	consistent
Oil temperature	−0.0002 [−0.0010, +0.0006]	−0.0079 [−0.0113, −0.0044]	−0.0077 [−0.0109, −0.0045]	consistent
Coolant temperature	+0.0000 [−0.0011, +0.0011]	−0.0009 [−0.0030, +0.0013]	−0.0009 [−0.0035, +0.0016]	unresolved

Only the named sensor is constrained; all other sensors are unconstrained. Reading: decision rules of Section 2.7.

3.6. Learned Directions from SHAP and EBM

SHAP and EBM agree with the audit (Table 10, Figure 7). For all three constrained sensors, both methods learned the physically expected directions. The contribution to the log-odds of a fault decreases with oil pressure (SHAP $\rho = -0.76$, $\Delta = -0.73$) and increases with oil temperature ($\rho = +0.96$, $\Delta = +0.79$) and coolant temperature ($\rho = +0.30$, $\Delta = +0.22$). The EBM without interaction terms gives the same main-effect shape functions (Figure A1): InterpretML estimates the main effects before the pairwise terms, so the interactions do not absorb them. Because SHAP and EBM are fitted to the same labels, their agreement

shows that the learned directions are robust to the model class; it is not independent evidence about the labels.

Coolant temperature is the weakest of the three. Its SHAP and EBM contributions rise steeply up to about 72 °C, are nearly flat between 72 and 86 °C, and fall again above about 87 °C (Figure 7). The decrease at the hottest end runs against overheating physics, but this region contains few records and the tail analysis there is inconclusive (Table 4). This weak and non-monotone effect explains why constraining coolant temperature alone gives an unresolved audit (Table 9).

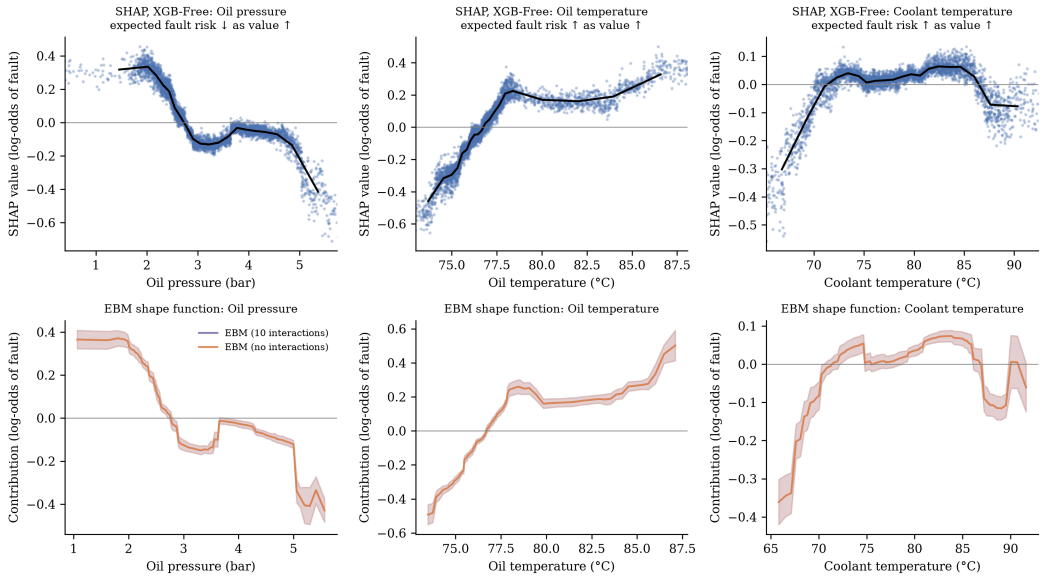


Figure 7. Learned effects of the three constrained sensors. Top: SHAP dependence of XGB-Free (points: test records; line: binned mean). Bottom: EBM shape functions with uncertainty bands, with and without interaction terms (the two curves coincide). The expected direction of fault risk is stated in each top panel title.

Table 10. Direction and size of learned effects on the test set, with bootstrap 95% confidence intervals.

Feature	Exp.	ρ SHAP [95% CI]	Δ SHAP [95% CI]	ρ EBM [95% CI]	Δ EBM [95% CI]	ρ EBM-0	Verdict
Engine speed	free	+0.99 [+0.99, +0.99]	+2.73 [+2.68, +2.78]	+1.00 [+1.00, +1.00]	+2.82 [+2.75, +2.87]	+1.00	n/a
Oil pressure	−	−0.76 [−0.78, −0.73]	−0.73 [−0.76, −0.70]	−0.76 [−0.78, −0.73]	−0.77 [−0.80, −0.73]	−0.76	Consistent
Fuel pressure	free	−0.95 [−0.96, −0.95]	−0.86 [−0.89, −0.83]	−0.98 [−0.99, −0.98]	−0.90 [−0.92, −0.88]	−0.98	n/a
Coolant pressure	free	+0.79 [+0.77, +0.81]	+0.21 [+0.18, +0.24]	+0.86 [+0.83, +0.88]	+0.20 [+0.17, +0.24]	+0.86	n/a
Oil temperature	+	+0.96 [+0.96, +0.97]	+0.79 [+0.76, +0.81]	+0.98 [+0.98, +0.98]	+0.82 [+0.80, +0.85]	+0.98	Consistent
Coolant temperature	+	+0.30 [+0.26, +0.34]	+0.22 [+0.20, +0.25]	+0.38 [+0.34, +0.42]	+0.22 [+0.19, +0.24]	+0.38	Consistent

Exp.: sign of the domain constraint; free: no directional expectation. Δ: Equation (6), change in log-odds of a fault from the lowest to the highest 5% of the feature. EBM-0: EBM without interaction terms. Consistent/Inconsistent require agreement of SHAP, EBM and EBM-0.

3.7. Semi-Synthetic Calibration

Table 11 and Figure 8 compare the real-data audit with the semi-synthetic scenarios. With consistent labels at the real effect size ($k = 1$), domain constraints were free (ΔAUC = +0.0000) and reversed constraints were costly (ΔAUC = −0.0138), giving PCG = −0.0139; the result was read as consistent. Inverting these labels produced the mirror image (PCG

= +0.0139, flagged). The symmetry follows by construction, because inverting all labels negates every learned effect.

The size of the signature grows with the strength of the constrained signal. For consistent labels, $\Delta AUC_{Reversed}$ was −0.0055, −0.0138, −0.0429 and −0.1203 for k = 0.5, 1, 2 and 4. The real reversed cost (−0.0159) lies between the values for k = 1 and k = 2. The real labels therefore behave like consistent labels of realistic strength. They are also compatible with a stronger consistent signal of which a minority of labels is randomly flipped: for k = 2, a reversed cost within 0.003 of the real value arises at q = 0.4. The audit cannot separate these two generating processes. What it can say is that most of the constrained signal follows the assumed physics.

Partial reversal gave intermediate values (PCG between −0.0122 and +0.0104). One of the 6 partial scenarios was flagged (reversed: oil pressure, oil temperature). One was nevertheless read as consistent (reversed: coolant temperature). A reversal confined to a sensor with a small coefficient can therefore go undetected, so a consistent reading means that most, not necessarily all, of the constrained signal follows the physics. Across the 30 null replicates, PCG ranged from −0.0352 to +0.0084. None was flagged, and none was read as consistent (Table 12, Figure 8c). The U-shaped rule N2 produced clearly negative PCG values, but its domain cost was significant, which is why the consistent reading requires a negligible domain cost. In the graded-flip series, however, the scenario with q = 0.5 and k = 1, whose labels carry no information (Free AUC = 0.486), was read as consistent (Table A3). The readings therefore presuppose an unconstrained model that discriminates better than chance, which should be checked first; the real labels (Free AUC = 0.703) meet this condition. The semi-synthetic experiment is deliberately simple and does not model the true data-generating process. Its purpose is to calibrate how the audit should be read.

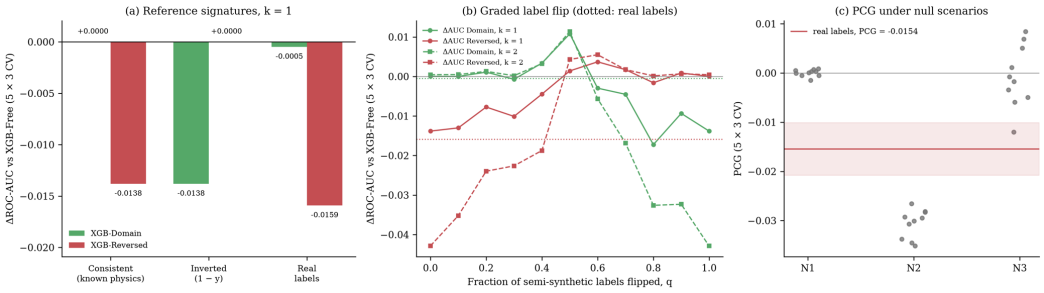


Figure 8. Semi-synthetic calibration (5 × 3 CV). (a) ΔROC-AUC relative to XGB-Free for consistent and inverted labels at the real effect size and for the real labels. (b) Graded label flip for two signal strengths; dotted lines mark the real-data values. (c) PCG in 10 replicates of each null scenario; the line and band show the real PCG and its 95% CI.

Table 11. Audit signatures under semi-synthetic labels with known physics and under the real labels (5 × 3 CV, ΔAUC relative to XGB-Free).

Scenario	Fault rate	Free AUC	ΔAUC Domain	ΔAUC Reversed	PCG [95% CI]	Reading
Consistent labels, k = 1	0.371	0.672	+0.0000	−0.0138	−0.0139 [−0.0182, −0.0095]	consistent
Inverted labels, k = 1	0.629	0.672	−0.0138	+0.0000	+0.0139 [+0.0095, +0.0182]	flagged
Consistent, k = 0.5	0.371	0.663	+0.0000	−0.0055	−0.0055 [−0.0083, −0.0026]	consistent
Inverted, k = 0.5	0.629	0.663	−0.0055	+0.0000	+0.0055 [+0.0026, +0.0083]	flagged
Consistent, k = 2	0.373	0.699	+0.0004	−0.0429	−0.0433 [−0.0506, −0.0361]	consistent
Inverted, k = 2	0.627	0.699	−0.0429	+0.0004	+0.0433 [+0.0360, +0.0506]	flagged
Consistent, k = 4	0.371	0.765	+0.0001	−0.1203	−0.1204 [−0.1303, −0.1104]	consistent
Inverted, k = 4	0.629	0.765	−0.1203	+0.0001	+0.1204 [+0.1104, +0.1303]	flagged
Reversed: Oil pressure	0.370	0.671	−0.0042	−0.0065	−0.0023 [−0.0087, +0.0041]	unresolved
Reversed: Oil temperature	0.371	0.668	−0.0054	−0.0075	−0.0021 [−0.0084, +0.0041]	unresolved
Reversed: Coolant temperature	0.370	0.671	−0.0008	−0.0130	−0.0122 [−0.0168, −0.0076]	consistent
Reversed: Oil temperature, Coolant temperature	0.372	0.667	−0.0062	−0.0058	+0.0004 [−0.0079, +0.0087]	unresolved

Scenario	Fault rate	Free AUC	ΔAUC Domain	ΔAUC Reversed	PCG [95% CI]	Reading
Reversed: Oil pressure, Coolant temperature	0.372	0.669	−0.0046	−0.0042	+0.0004 [−0.0062, +0.0069]	unresolved
Reversed: Oil pressure, Oil temperature	0.375	0.668	−0.0104	+0.0000	+0.0104 [+0.0044, +0.0165]	flagged
Real dataset labels	0.369	0.703	−0.0005	−0.0159	−0.0154 [−0.0208, −0.0101]	consistent

Partial reversal: the physical sign is reversed only for the named sensors. Reading: decision rules of Section 2.7. Graded-flip scenarios are listed in Table A3.

Table 12. PCG under null scenarios (10 replicates each, 5 × 3 CV).

Scenario	Replicates	Mean PCG	Range	Share flagged	Share consistent
N1 unconstrained sensors only	10	+0.0000	[−0.0015, +0.0009]	0.0	0.0
N2 non-monotone (U-shaped) rule	10	−0.0306	[−0.0352, −0.0266]	0.0	0.0
N3 permuted real labels	10	−0.0007	[−0.0120, +0.0084]	0.0	0.0
Real dataset labels	1	−0.0154	—	0.0	1.0

N1: labels depend only on the unconstrained sensors; N2: non-physical U-shaped rule on the constrained sensors; N3: randomly permuted real labels.

3.8. Positive Control on the AI4I 2020 Dataset

On the AI4I 2020 dataset, whose failure mechanisms are documented, the audit behaved as expected (Table 13). With the documented labels, domain constraints were almost free (ΔAUC = +0.0001) and reversed constraints were costly (ΔAUC = −0.0909), giving PCG = −0.0910 ([−0.1254, −0.0566]); the result was read as consistent. With inverted labels, the gap changed sign (PCG = +0.0910, [+0.0566, +0.1254]) and the dataset was flagged. The hold-out test gave the same picture. The audit thus recognises correct physics as correct and inverted labels as inverted on a dataset where the truth is known.

Table 13. Positive control: audit on the AI4I 2020 dataset with documented and inverted labels.

Labels	Free AUC	ΔAUC Domain [95% CI]	ΔAUC Reversed [95% CI]	CV PCG [95% CI]	Test PCG [95% CI]	Reading
Documented labels	0.978	+0.0001 [−0.0014, +0.0017]	−0.0909 [−0.1258, −0.0560]	−0.0910 [−0.1254, −0.0566]	−0.0711 [−0.1092, −0.0395]	consistent
Inverted labels (1 − y)	0.978	−0.0909 [−0.1258, −0.0560]	+0.0001 [−0.0014, +0.0017]	+0.0910 [+0.0566, +0.1254]	+0.0711 [+0.0395, +0.1092]	flagged

Constraints: temperature difference −1, tool wear +1; product type, air temperature, speed and torque unconstrained. 5 × 3 CV on the training part; test: paired bootstrap on the 20% hold-out part.

3.9. Implications for the Label Convention and the Data

Under the documented coding, the audit reads the labels as consistent with the assumed physics, and the semi-synthetic calibration, the null scenarios and the positive control show how this reading should be interpreted. Inverting all labels negates every learned effect and leaves ROC-AUC unchanged, so the opposite coding exchanges the roles of the two constraint sets. Running the complete protocol with Engine Condition = 1 as the fault class confirmed this: ΔAUC_{Domain} = −0.0142, $\Delta AUC_{Reversed}$ = −0.0007 and PCG = +0.0136 ([+0.0090, +0.0181]), which is flagged; the small differences from the mirror image of the present results come from re-tuning. This is the coding of the first version of this study. The repository record does not define the coding, and a user question on the record asking which value means good has no answer there [25]. A wrong convention is therefore an easy mistake, and it is invisible to accuracy, calibration and label-noise detectors [22], because swapping the classes leaves model confidence unchanged. The audit detected it from directional knowledge alone. This is the most practical use of the method shown here: a quick check of the label convention of a poorly documented dataset.

Three observations limit what the consistent reading means. First, the tail analysis is mixed: some extremes agree with physics, others contradict it, and the low-oil-pressure tail is uninformative. Second, the data are synthetic (Section 2.1). The fault class derives from three genuine fault observations, oversampled and expanded by a generator, and

the near-zero inter-sensor correlations show that the sensors were not generated as a coupled engine system. The consistent reading therefore shows that the generated labels preserve the expected dominant directions of oil pressure and oil temperature; it does not show that they describe real engine failures. Third, directional agreement can also arise from confounding, from sensors that are missing from the file, or from monotone assumptions that are too simple. The audit tests agreement between labels and assumed physics; it cannot by itself establish that the labels are correct.

3.10. Practical Implications and Limitations

For practitioners, the audit is inexpensive. It requires two additional model fits and a short list of directional expectations that domain experts can readily supply, and it should be run before a detector is trusted, in particular when the label definition is undocumented or taken from a secondary source. A flag should trigger a review of the label coding and of the data-collection or simulation process. Models trained on the present synthetic data, including EBM, should not be deployed on real engines; EBM is recommended here as an audit instrument, because it exposes its learned effects directly to engineers.

Several limitations apply. First, the analysis uses one synthetic maritime dataset and one positive control, and the achievable discrimination is modest (ROC-AUC ≈ 0.70). Second, the benchmark tuning is not nested; the hold-out results and the nested audit mitigate but do not remove this issue. Third, manufacturer alarm set points were not available, so the tail regions are conservative extremes. Fourth, monotonic constraints cannot express non-monotone risk such as that of engine speed, fuel pressure or very low coolant temperature; shape constraints beyond monotonicity would be needed. Fifth, the file contains independent snapshots without machine identifiers or timestamps, which restricts the task to fault detection rather than prognosis. Future work should apply the audit to measured datasets with documented labels and run-to-failure histories and extend it to non-monotone domain knowledge.

4. Conclusions

This study proposed and evaluated a physical-consistency audit for engine fault detection that combines domain-informed monotonic XGBoost with the glass-box Explainable Boosting Machine. On a public marine diesel engine dataset, the best classifiers reached a modest test ROC-AUC of about 0.71. Addressing the first research gap, the three-way audit showed that, under the label coding documented by the dataset creator, engineering-consistent constraints had no measurable cost, whereas reversed constraints reduced the test ROC-AUC by 0.015 and removed almost all reliance on oil pressure and oil temperature. The resulting physical-consistency gap (PCG = −0.0154) was robust to separate and nested tuning, and SHAP and EBM agreed on the expected directions. Addressing the second research gap, semi-synthetic calibration, null scenarios and a positive control showed how the gap should be read, and the opposite label coding, used in the first version of this study, was flagged. The audit thus detects a wrong label convention, which accuracy-based evaluation cannot. Mixed tail evidence and the synthetic origin of the data show that consistent directions do not by themselves make labels valid for real engines. Physical-consistency auditing should therefore complement accuracy-based evaluation, and label definitions should be documented and checked before fault-detection models are trusted in maintenance practice.

Supplementary Materials: The Python code of the revised experiment (audit_experiment_v3.py), the result tables (results_v3.xlsx), the summary file (summary_v3.json), the package versions (environment.txt) and all figures are provided as supplementary files and are archived on Zenodo (https://doi.org/10.5281/zenodo.23728914).

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft preparation, writing—review and editing, and visualization, A.A.A. The author has read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The ship main-engine dataset is publicly available on IEEE DataPort (https://doi.org/10.21227/g3za-v415) [25]; the file engine_data.csv was downloaded on 29 June 2026 (MD5 629ad357fe0e5d0e6beb7421e2ba2c32). The AI4I 2020 dataset is available from the UCI Machine Learning Repository [32]. The code needed to reproduce all results is provided in the Supplementary Materials and on Zenodo (https://doi.org/10.5281/zenodo.23728914).

Acknowledgments: The author thanks the providers of the public engine-condition dataset.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Table A1. Hyper-parameters selected by randomised search with three-fold CV on the training set.

Model	Tuned hyper-parameters	Inner-CV ROC-AUC
Logistic Regression	C: 0.006158482110660267	0.6876
Random Forest	min_samples_leaf: 10, max_features: sqrt, max_depth: 6	0.7000
LightGBM	subsample: 0.85, num_leaves: 7, n_estimators: 200, min_child_samples: 50, learning_rate: 0.05, colsample_bytree: 1.0	0.7025
CatBoost	learning_rate: 0.05, l2_leaf_reg: 10, iterations: 200, depth: 3	0.7047
XGBoost (XGB-Free; shared by the main audit)	subsample: 0.7, n_estimators: 200, min_child_weight: 20, max_depth: 4, learning_rate: 0.03, colsample_bytree: 0.85	0.7031
EBM	max_bins: 128, learning_rate: 0.03, interactions: 10	0.7057
XGB-Domain (own tuning)	subsample: 0.85, n_estimators: 500, min_child_weight: 10, max_depth: 4, learning_rate: 0.01, colsample_bytree: 0.7	0.7029
XGB-Reversed (own tuning)	subsample: 1.0, n_estimators: 100, min_child_weight: 20, max_depth: 3, learning_rate: 0.05, colsample_bytree: 0.85	0.6888
XGBoost (AI4I 2020 positive control)	subsample: 0.7, n_estimators: 300, min_child_weight: 1, max_depth: 4, learning_rate: 0.01, colsample_bytree: 1.0	—

EBM without interactions uses the EBM hyper-parameters with interactions = 0.

Table A2. Sensitivity of the audit to the cleaning rule (5 × 3 CV on the respective training partition).

Variant	Records	ΔAUC Domain	ΔAUC Reversed	PCG [95% CI]	Reading
Main analysis (coolant temperature > 110 °C removed)	19,533	−0.0005	−0.0159	−0.0154 [−0.0208, −0.0101]	consistent
No record removed	19,535	−0.0005	−0.0155	−0.0151 [−0.0209, −0.0093]	consistent
Only the 195.5 °C record removed	19,534	−0.0004	−0.0161	−0.0156 [−0.0201, −0.0112]	consistent
Also oil pressure < 0.5 bar removed	19,415	−0.0007	−0.0164	−0.0156 [−0.0202, −0.0110]	consistent

Table A3. Graded label flip: audit signatures when a fraction q of consistent semi-synthetic labels is inverted.

k	q	ΔAUC Domain	ΔAUC Reversed	PCG [95% CI]	Reading
1	0.0	+0.0000	−0.0138	−0.0139 [−0.0182, −0.0095]	consistent
1	0.1	+0.0000	−0.0130	−0.0130 [−0.0171, −0.0090]	consistent
1	0.2	+0.0011	−0.0077	−0.0088 [−0.0151, −0.0025]	consistent
1	0.3	−0.0007	−0.0101	−0.0094 [−0.0186, −0.0002]	consistent
1	0.4	+0.0034	−0.0045	−0.0078 [−0.0212, +0.0055]	unresolved
1	0.5	+0.0108	+0.0014	−0.0094 [−0.0162, −0.0025]	consistent
1	0.6	−0.0030	+0.0037	+0.0067 [−0.0047, +0.0180]	unresolved
1	0.7	−0.0045	+0.0017	+0.0062 [−0.0011, +0.0136]	unresolved
1	0.8	−0.0172	−0.0016	+0.0157 [+0.0057, +0.0257]	flagged
1	0.9	−0.0093	+0.0009	+0.0102 [+0.0023, +0.0181]	flagged
1	1.0	−0.0138	+0.0000	+0.0139 [+0.0095, +0.0182]	flagged
2	0.0	+0.0004	−0.0429	−0.0433 [−0.0506, −0.0361]	consistent
2	0.1	+0.0005	−0.0352	−0.0357 [−0.0425, −0.0289]	consistent
2	0.2	+0.0013	−0.0240	−0.0253 [−0.0341, −0.0164]	consistent
2	0.3	+0.0002	−0.0226	−0.0228 [−0.0316, −0.0140]	consistent

k	*q*	ΔAUC Domain	ΔAUC Reversed	PCG [95% CI]	Reading
2	0.4	+0.0032	−0.0187	−0.0220 [−0.0301, −0.0139]	consistent
2	0.5	+0.0114	+0.0043	−0.0070 [−0.0214, +0.0073]	unresolved
2	0.6	−0.0056	+0.0055	+0.0111 [+0.0022, +0.0200]	flagged
2	0.7	−0.0168	+0.0017	+0.0186 [+0.0093, +0.0279]	flagged
2	0.8	−0.0326	+0.0002	+0.0327 [+0.0246, +0.0409]	flagged
2	0.9	−0.0323	+0.0007	+0.0330 [+0.0247, +0.0413]	flagged
2	1.0	−0.0429	+0.0004	+0.0433 [+0.0360, +0.0506]	flagged

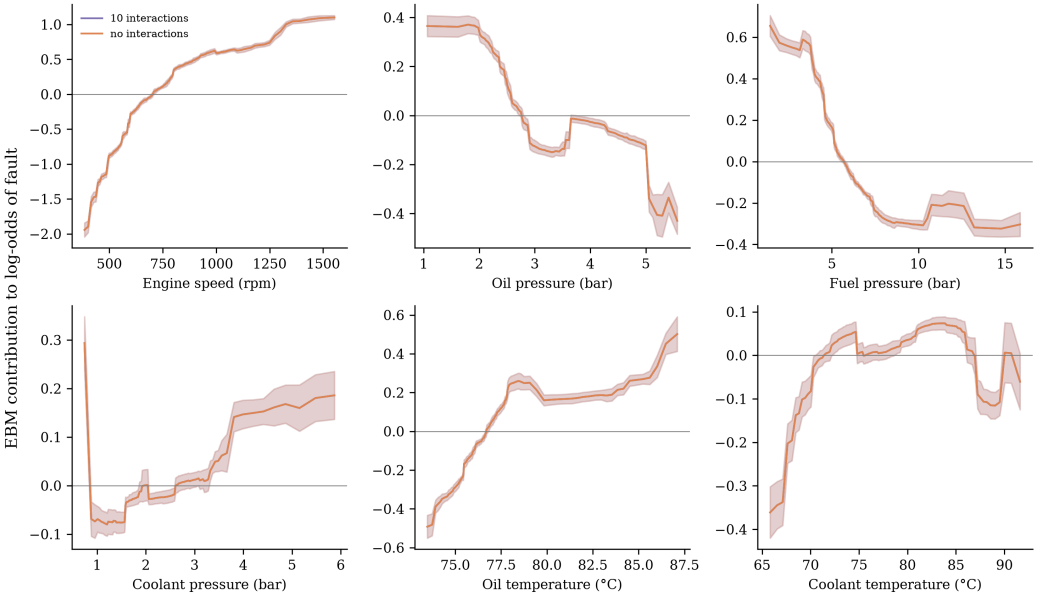


Figure A1. EBM shape functions for all six sensor variables (contribution to the log-odds of a fault), with and without interaction terms, trimmed to the 0.5–99.5th percentile range.

References

1. Jardine, A.K.S.; Lin, D.; Banjevic, D. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mech. Syst. Signal Process.* **2006**, *20*, 1483–1510.
2. Lei, Y.; Li, N.; Guo, L.; Li, N.; Yan, T.; Lin, J. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. *Mech. Syst. Signal Process.* **2018**, *104*, 799–834.
3. Mohakul, D.; Kumar, C.R.S.; Singh, S.; Katti, S. Condition based predictive maintenance on ship's major equipment using AI. *Int. J. Sci. Res. Eng. Manag.* **2023**, *7*(2), 1–8. https://doi.org/10.55041/IJSREM17699.
4. Carvalho, T.P.; Soares, F.A.A.M.N.; Vita, R.; Francisco, R.P.; Basto, J.P.; Alcalá, S.G.S. A systematic literature review of machine learning methods applied to predictive maintenance. *Comput. Ind. Eng.* **2019**, *137*, 106024.
5. Zhang, W.; Yang, D.; Wang, H. Data-driven methods for predictive maintenance of industrial equipment: A survey. *IEEE Syst. J.* **2019**, *13*, 2213–2227.
6. Cummins, L.; Sommers, A.; Ramezani, S.B.; Mittal, S.; Jabour, J.; Seale, M.; Rahimi, S. Explainable predictive maintenance: A survey of current methods, challenges and opportunities. *IEEE Access* **2024**, *12*, 57574–57602. https://doi.org/10.1109/ACCESS.2024.3391130.
7. Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 1721–1730.
8. Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; Wichmann, F.A. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2020**, *2*, 665–673.
9. Lapuschkin, S.; Wäldchen, S.; Binder, A.; Montavon, G.; Samek, W.; Müller, K.-R. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.* **2019**, *10*, 1096.
10. Ross, A.S.; Hughes, M.C.; Doshi-Velez, F. Right for the right reasons: Training differentiable models by constraining their explanations. In Proceedings of the 26th International Joint

Conference on Artificial Intelligence (IJCAI-17), Melbourne, Australia, 19–25 August 2017; pp. 2662–2670. https://doi.org/10.24963/ijcai.2017/371.
11. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215.
12. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
13. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67.
14. Karpatne, A.; Atluri, G.; Faghmous, J.H.; Steinbach, M.; Banerjee, A.; Ganguly, A.; Shekhar, S.; Samatova, N.; Kumar, V. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 2318–2331.
15. Karniadakis, G.E.; Kevrekidis, I.G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-informed machine learning. *Nat. Rev. Phys.* **2021**, *3*, 422–440.
16. Gupta, M.; Cotter, A.; Pfeifer, J.; Voevodski, K.; Canini, K.; Mangylov, A.; Moczydlowski, W.; van Esbroeck, A. Monotonic calibrated interpolated look-up tables. *J. Mach. Learn. Res.* **2016**, *17*, 1–47.
17. Koklev, P. What's the price of monotonicity? A multi-dataset benchmark of monotone-constrained gradient boosting for credit PD. *arXiv* **2025**, arXiv:2512.17945.
18. Abbas, S.; Thomas, D.; Soomro, N.; Shafik, R.; Heer, R.; Adhikari, K. Fixing a model that learned worse cancer means lower risk: Monotonic constraints in bladder cancer recurrence prediction. *arXiv* **2026**, arXiv:2610.00858.
19. Hsiao, C.-H.; Kumar, K.; Rathje, E.M. Domain-informed explainable boosting machines for trustworthy lateral spread predictions. *arXiv* **2026**, arXiv:2603.17175.
20. Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J.W.; Wallach, H.; Daumé III, H.; Crawford, K. Datasheets for datasets. *Commun. ACM* **2021**, *64*, 86–92.
21. Frénay, B.; Verleysen, M. Classification in the presence of label noise: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2014**, *25*, 845–869.
22. Northcutt, C.; Jiang, L.; Chuang, I. Confident learning: Estimating uncertainty in dataset labels. *J. Artif. Intell. Res.* **2021**, *70*, 1373–1411.
23. Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O. Leakage in data mining: Formulation, detection, and avoidance. *ACM Trans. Knowl. Discov. Data* **2012**, *6*, 15.
24. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804.
25. Mohakul, D. Predictive Maintenance on Ship's Main Engine using AI [Dataset]. IEEE DataPort, 16 November 2022. https://doi.org/10.21227/g3za-v415 (file engine_data.csv, accessed on 29 June 2026).
26. Nurdin, F.; Effendi, M.K.; Mohakul, D. Machine learning-based prediction of diesel engine health using operational parameters: Comparison of SVM and BPNN models. *Int. J. Mech. Eng. Sci.* **2025**, *9*, 69–80. https://doi.org/10.12962/j25807471.v9i2.10511.
27. Chukwudi, I.J.; Zaman, N.; Rahim, M.A.; Rahman, M.A.; Alenazi, M.J.F.; Pillai, P. An ensemble deep learning model for vehicular engine health prediction. *IEEE Access* **2024**, *12*, 63433–63451. https://doi.org/10.1109/ACCESS.2024.3395927.
28. Harrou, F.; Dorbane, A.; Sun, Y. Explainable machine learning with SMOTE for predictive maintenance of ship engines. *Proc. Inst. Mech. Eng. Part M J. Eng. Marit. Environ.* **2025**, online ahead of print. https://doi.org/10.1177/14750902251400373.
29. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
30. Lou, Y.; Caruana, R.; Gehrke, J.; Hooker, G. Accurate intelligible models with pairwise interactions. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11–14 August 2013; pp. 623–631.
31. Nori, H.; Jenkins, S.; Koch, P.; Caruana, R. InterpretML: A unified framework for machine learning interpretability. *arXiv* **2019**, arXiv:1909.09223.
32. Matzka, S. Explainable artificial intelligence for predictive maintenance applications. In Proceedings of the 2020 Third International Conference on Artificial Intelligence for Industries (AI4I), Irvine, CA, USA, 21–23 September 2020; pp. 69–74. https://doi.org/10.1109/AI4I49448.2020.00023.

33. Patki, N.; Wedge, R.; Veeramachaneni, K. The Synthetic Data Vault. In Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, 17–19 October 2016; pp. 399–410. https://doi.org/10.1109/DSAA.2016.49.
34. Mohakul, D.; Kumar, C.; Singh, S.; Katti, S.; Chougule, S. Health monitoring of ship's engine with simulated data by using classifiers—preliminary result. In Proceedings of the 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichirappalli, India, 5–7 April 2023; pp. 1–5.
35. Varma, S.; Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **2006**, *7*, 91.
36. Youden, W.J. Index for rating diagnostic tests. *Cancer* **1950**, *3*, 32–35.
37. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
38. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
39. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 3146–3154.
40. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, QC, Canada, 3–8 December 2018; pp. 6638–6648.
41. Woodyard, D. *Pounder's Marine Diesel Engines and Gas Turbines*, 9th ed.; Butterworth-Heinemann: Oxford, UK, 2009.
42. Nadeau, C.; Bengio, Y. Inference for the generalization error. *Mach. Learn.* **2003**, *52*, 239–281.
43. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman & Hall: New York, NY, USA, 1993.
44. Wilson, E.B. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* **1927**, *22*, 209–212.
45. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432.
46. Van Calster, B.; McLernon, D.J.; van Smeden, M.; Wynants, L.; Steyerberg, E.W. Calibration: The Achilles heel of predictive analytics. *BMC Med.* **2019**, *17*, 230.
47. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6.
48. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
49. Schuirmann, D.J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J. Pharmacokinet. Biopharm.* **1987**, *15*, 657–680.
50. Benavoli, A.; Corani, G.; Demšar, J.; Zaffalon, M. Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis. *J. Mach. Learn. Res.* **2017**, *18*, 1–36.