International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(23s) 2026 pp. 1088-1117
https://doi.org/10.70917/ijcisim-2026-5682
Article

Modernizing Legacy Safety Instrumented Systems: A SIL 2 Predictive Gas Detection Framework for Global IEC 61511 Compliance and Industrial Risk Reduction

Nazim Nazir[1], Anuradha Kalansuriya[2], Seemab Javed[3], Zunaira Irfan[4], Rashed Abu Hammour[5]

[1] University of Engineering and Technology PK Lahore. Email: Nnk_2012@hotmail.com ORCID: 0009-0007-4597-8896
[2] Department of Engineering, Technology & Quantity Surveying, INTI International University, Nilai, Negeri Sembilan, Malaysia. Email: i25033404@student.newinti.edu.my ORCID: 0009-0005-9587-8338
[3] University of engineering and technology Lahore. Email: Seemab.rajput@gmail.com ORCID: 0009-0008-3409-6975
[4] Department of Chemical Engineering, UET. Email: zanairairfan@gmail.com
[5] Faculty of Technical Education, Hourani Center for Applied Scientific Research (HCASR), Al-Ahliyya Amman University, Amman, Jordan. Email: r.abuhammour@ammanu.edu.jo
Corresponding Author: Seemab Javed

Abstract: Ageing safety instrumented systems (SIS) remain the dominant protective layer against flammable and toxic gas releases in the process industries, yet a large installed base of legacy gas detection loops was engineered before IEC 61511 and cannot demonstrate the safety integrity level (SIL) that current layer of protection analysis (LOPA) assigns to them. This study develops and evaluates a Predictive Gas Detection Framework (PGDF), a four-layer retrofit architecture that couples hardware modernization with a supervised prognostic layer that forecasts dangerous undetected (DU) detector failures before they occur. Using a fully specified simulation study of a brownfield gas processing facility, we (a) verified the average probability of failure on demand (PFDavg) of a legacy 1oo2 catalytic-bead gas detection safety instrumented function (SIF) and of two modernization options by 120,000-iteration Monte Carlo analysis with common random numbers; (b) trained and temporally validated four classifiers on 20,493 weekly detector-health records from 184 simulated detectors; and (c) propagated the resulting diagnostic performance through a 36-month deployment simulation, a LOPA, and a 15-year cost model. The legacy SIF achieved a median PFDavg of 1.73×10^{-2} (90% credible interval [CrI] 1.01×10^{-2}–2.99×10^{-2}; risk reduction factor [RRF] 58), meeting SIL 2 in only 4.6% of iterations. Like-for-like sensor and logic-solver replacement was insufficient (PFDavg 1.40×10^{-2}; SIL 2 in 15.6%). The full PGDF retrofit achieved 1.88×10^{-3} (90% CrI 9.88×10^{-4}–3.64×10^{-3}; RRF 533), satisfying SIL 2 in 100.0% of iterations, a paired reduction of 89.0% (95% CI 82.4–93.9). A random forest was the strongest prognostic model (area under the receiver operating characteristic curve [AUC] 0.849, 95% CI 0.836–0.862), significantly exceeding elastic-net logistic regression (ΔAUC 0.024, $p < .001$). Discrimination was excellent for gradual degradation failures (AUC 0.892) but no better than chance for abrupt shock-induced failures (AUC 0.545), which imposes a hard ceiling on any prognostic layer. Mean covert-failure exposure fell from 88.2 to 21.1 days (−76.1%; $t = 11.06$, $p < .001$; Hedges' $g = 1.34$, 95% CI 1.08–1.61), 75.7% of DU failures were averted before onset (95% CI 67.6–82.7), and the spurious alarm rate fell by 69.0% (incidence rate ratio 0.31, 95% CI 0.26–0.38, $p < .001$). Only the full PGDF met the tolerable event frequency of 2×10^{-5} yr^{-1} (mitigated frequency 5.63×10^{-6} yr^{-1}). Discounted payback was 4.68 years. We conclude that predictive diagnostics are a legitimate but bounded contributor to SIL attainment: they raise effective diagnostic coverage for degradation-mode failures and shorten covert exposure, but hardware redundancy and final-element integrity remain the load-bearing elements of compliance. Findings are derived from simulated data and require field validation before operational adoption.

Keywords: Safety instrumented system; IEC 61511; safety integrity level; gas detection; predictive maintenance; probability of failure on demand; layer of protection analysis; process safety; machine learning; functional safety

1. Introduction

Fixed gas detection is the most widely deployed safety instrumented function in the hydrocarbon processing, petrochemical and liquefied natural gas sectors. It is the protective layer that converts an unignited loss of containment into a controlled shutdown rather than a fire, an explosion or a toxic exposure. Because a gas detection safety instrumented function (SIF) operates in low-demand mode, its failures are overwhelmingly covert: a poisoned catalytic bead, a fouled optical path or a drifted electrochemical cell continues to output a plausible 4–20 mA signal while providing no protection at all. The safety burden therefore falls not on the detector's nominal accuracy but on the plant's ability to discover that the detector has stopped working before a demand arrives. Recent work on data-driven leak detection has shown that machine learning can extract early release signatures from spatial and temporal sensor patterns (Kopbayev et al., 2022) and from engineered features of natural gas leakage data (Kang et al., 2023), but this literature has concentrated almost entirely on detecting the process event rather than on detecting the failure of the instrument that is supposed to see it.

The International Electrotechnical Commission (IEC) standard 61511 governs the entire safety lifecycle of such systems in the process sector, translating the generic requirements of IEC 61508 into process-industry terminology and establishing the safety integrity level (SIL) as the order-of-magnitude measure of required risk reduction (International Electrotechnical Commission [IEC], 2016a, 2016b). A SIF is credited with a SIL only if three independent barriers are cleared simultaneously: the quantified average probability of failure on demand (PFDavg) must fall within the SIL band, the architectural constraints on hardware fault tolerance and safe failure fraction must be satisfied, and the systematic capability of every element must be demonstrated. A great deal of industrial practice treats the first barrier as if it were the whole test, which produces the recurring finding that installed SIFs are assigned a SIL in the design documentation that the as-operated hardware cannot support.

The scale of the legacy problem is structural rather than incidental. A substantial proportion of gas detection loops now in service were engineered under the first edition of ANSI/ISA S84 or under no functional safety standard at all, using relay logic or first-generation programmable logic controllers whose diagnostic coverage is low and whose systematic capability cannot be evidenced retrospectively. These loops are then re-baselined against a modern LOPA that assigns them SIL 2. The resulting compliance gap is rarely closed, because the obvious remedy — shortening the proof-test interval — consumes maintenance capacity that brownfield facilities do not have, and because full replacement is capital-intensive. Analytical treatments of proof-test interval and coverage confirm that PFDavg is approximately proportional to the test interval only for the fraction of dangerous undetected failures that the test actually reveals, so that imperfect testing places a floor beneath the achievable PFDavg regardless of test frequency (Gu et al., 2022).

A parallel body of work has established that the health state of industrial sensors is itself learnable. Drift in metal-oxide and catalytic sensing elements follows characteristic trajectories that can be compensated or predicted using recurrent architectures (Chaudhuri et al., 2021; Zhuang et al., 2024), reviewed systematically for selectivity in metal-oxide devices by Djeziri et al. (2024). Convolutional autoencoders have been used for real-time sensor fault detection, localization and correction in instrumented structures (Jana et al., 2022), gradient-boosted ensembles for vibration sensor fault diagnosis (Fan et al., 2023), and unsupervised multivariate anomaly detection for industrial internet-of-things deployments (Belay et al., 2023). Dynamic reliability digital twins driven by artificial intelligence have been shown to support predictive maintenance of standalone industrial components (D'Urso et al., 2024). What is missing is the connective step: none of these contributions expresses its diagnostic gain in the currency that a functional safety assessment actually consumes, which is a change in dangerous undetected failure rate, in effective diagnostic coverage, and hence in PFDavg and SIL.

This omission matters because the conversion is not automatic and is easy to overstate. A prognostic model that flags a degrading detector reduces the time that detector spends in an undetected failed state; it does not reduce the rate at which detectors fail, and it cannot address failure modes that produce no precursor. Abrupt failures — water ingress into a junction box, mechanical impact, a terminal or cable fault — arrive without a degradation trajectory to

learn from. Any honest claim about predictive maintenance and SIL must therefore partition the failure population by mechanism and quantify performance separately within each partition. Recent reviews of assurance for safety-critical artificial intelligence in the process industries make the same point from the governance side: a data-driven component can fail silently and return confident predictions in operating regimes that are under-represented in its training data, which is precisely the behaviour that the systematic capability and operational-verification requirements of IEC 61511 exist to guard against (IEC, 2016a).

The present study addresses that gap directly. We define a Predictive Gas Detection Framework (PGDF), a four-layer retrofit architecture for legacy gas detection SIFs, and we evaluate it end to end within a single internally consistent simulation. The evaluation is deliberately structured so that every claim is traceable to the quantity that IEC 61511 recognizes. Specifically, we ask four questions. First, what PFDavg and SIL does a representative legacy gas detection SIF actually achieve once parameter uncertainty is propagated, and is the widely used remedy of like-for-like sensor replacement sufficient? Second, how accurately can supervised models forecast dangerous undetected detector failures from routinely available health telemetry, and does that accuracy hold across detector technologies and failure mechanisms? Third, when the prognostic layer is embedded in a realistic maintenance workflow, how much covert-failure exposure, spurious alarm burden and proof-test labour does it actually remove? Fourth, does the modernized architecture close the LOPA gap, and at what cost?

Two commitments shape the way results are reported. First, we distinguish throughout between mechanisms the analysis demonstrates within its own assumptions and propositions that remain hypotheses requiring field data, and we tabulate that distinction explicitly. Second, we resist the framing in which predictive analytics substitute for hardware integrity. The central empirical finding of this work is that they do not: the prognostic layer contributes a real but bounded share of the achieved risk reduction, and the majority is contributed by redundancy and final-element integrity. Presenting that partition honestly is more useful to a functional safety assessor than a larger undifferentiated improvement figure would be.

2. Materials and Methods

2.1 Study design and reporting

This is a modelling and simulation study. It comprises three coupled components: a probabilistic reliability analysis of three SIS architectures, a supervised learning experiment on detector health telemetry, and a discrete deployment simulation that propagates the learned diagnostic performance into operational, risk and economic outcomes. No field data from an operating facility were used. All data are synthetic, generated from reliability parameter ranges consistent with published industrial reliability databases and from a physically motivated stochastic degradation model specified in full in Section 2.6. This design choice is a deliberate trade: it permits complete knowledge of the ground-truth failure state, exact control of the failure-mechanism mixture, and full reproducibility, at the cost of external validity. The consequences are addressed in Section 4.10. Every numerical result reported in Section 3 is the output of a single deterministic analysis script executed with a fixed pseudo-random seed.

2.2 Reference facility and the target safety instrumented function

The reference case is a brownfield onshore gas processing train commissioned in the early 1990s and re-assessed under a contemporary LOPA. The target SIF, designated SIF-101, detects a flammable hydrocarbon release in the compressor house and initiates emergency shutdown and isolation of the compressor suction. Table 1 specifies the function, its demand characteristics and its as-installed architecture.

Table 1

Specification of the target safety instrumented function (SIF-101)

Attribute	Specification
Safety instrumented function	SIF-101 — flammable gas detection and compressor suction isolation

Hazardous event	Ignited hydrocarbon release in an enclosed compressor house
Detected variable	Methane / C_1–C_4 hydrocarbons, 0–100% lower explosive limit (LEL)
Alarm and trip set points	Low alarm 10% LEL; executive action 20% LEL (coincidence voting)
Demand mode	Low demand (< 1 demand yr^{-1})
Demand rate on the SIF	3.0×10^{-3} yr^{-1} (after credit for non-SIS protection layers)
Target safety integrity level	SIL 2 (required risk reduction factor 150)
As-installed sensor subsystem	1oo2 catalytic bead detectors, 12-month proof-test interval
As-installed logic solver	Relay logic with first-generation PLC supervision, HFT = 0
As-installed final element	Solenoid-operated emergency shutdown valve, 1oo1, no partial stroke testing
Site detector population	186 fixed detectors across six process units
Assumed mission time	20 years

2.3 The Predictive Gas Detection Framework

PGDF is specified as four layers, each of which maps to a distinct clause family in IEC 61511-1. The layering is deliberate: it keeps the machine-learning component outside the safety-rated execution path, so that a prognostic failure degrades maintenance efficiency but cannot defeat the safety function. Layer 1 replaces the sensing and logic hardware. Layer 2 acquires and conditions non-safety health telemetry from the same devices over a read-only interface. Layer 3 is the supervised prognostic engine. Layer 4 converts prognostic output into scheduled maintenance action and into revised proof-test planning. Table 2 gives the full specification and Figure 1 shows the information flow.

Table 2

Specification of the four PGDF layers and their functional safety role

Layer	Elements	Function	Safety classification
1. Safety-rated hardware	2oo3 infrared point detectors; open-path detectors at unit boundaries; SIL 3 certified safety PLC; 1oo2 emergency shutdown valves with quarterly partial stroke testing	Executes the safety instrumented function; provides hardware fault tolerance and on-line automatic diagnostics	Safety-related; within the SIF boundary
2. Telemetry acquisition	Edge gateway; read-only HART or Modbus interface; historian with weekly aggregation of span response, T_{90}, baseline offset, bump-test residual and supply-current noise	Extracts device health variables without writing to the safety loop	Non-interfering; strict one-way data diode
3. Prognostic engine	Feature engineering (levels, 4- and 12-week velocities, exponentially weighted means); supervised classifier; calibrated probability of a dangerous undetected failure within 90 days	Forecasts covert failure; ranks the detector population by residual integrity	Non-safety; advisory only
4. Maintenance and assurance	Risk-ranked work-order generation; condition-based intervention; evidence package for the functional safety assessment	Converts forecast into intervention; documents diagnostic coverage claims	Procedural; subject to management of change

PROCESS DEMAND → Flammable hydrocarbon release in the compressor house		
LAYER 1 — SAFETY-RATED EXECUTION PATH (inside the SIF boundary)		
2oo3 IR point detectors + open-path boundary detectors	SIL 3 certified safety PLC	1oo2 ESD valves + quarterly partial stroke test
↓ *read-only telemetry tap (one-way; no write path back to Layer 1)* ↓		
LAYER 2 — TELEMETRY ACQUISITION (outside the SIF boundary)		
Edge gateway HART / Modbus read	Historian weekly aggregation	Span response, T_{90}, baseline offset, bump residual, supply-current noise
↓		
LAYER 3 — PROGNOSTIC ENGINE (advisory only)		
Feature engineering levels, 4- and 12-week velocities, EWMA	Supervised classifier random forest	Calibrated P(DU failure within 90 days)
↓		
LAYER 4 — MAINTENANCE AND ASSURANCE (procedural)		
Risk-ranked work orders	Condition-based intervention	Revised proof-test plan + FSA evidence package
→ *feeds back to Layer 1 as physical maintenance action, never as a control signal* →		

Figure 1. Information flow through the four PGDF layers. Only Layer 1 lies inside the safety instrumented function boundary. The telemetry tap is strictly read-only, and Layer 4 closes the loop through physical maintenance action rather than through any signal path into the safety-rated execution path, so that a prognostic failure degrades maintenance efficiency without defeating the protective function. IR = infrared; PLC = programmable logic controller; ESD = emergency shutdown; EWMA = exponentially weighted moving average; DU = dangerous undetected; FSA = functional safety assessment.

2.4 Reliability modelling and Monte Carlo uncertainty propagation

PFDavg was computed for three architectures: the as-installed baseline; an interim option representing the common industrial response of replacing sensors and the logic solver while retaining calendar-based testing and the original single shutdown valve; and the full PGDF retrofit. Subsystem PFDavg was evaluated using the standard low-demand approximations of IEC 61508-6 with an explicit β-factor treatment of common cause failure and an additive test-independent failure term. For a single (1oo1) element,

$$PFD_{1oo1} = \lambda_{DU} \cdot TI / 2 + \lambda_{DD} \cdot MTTR + P_T^{If} \quad (1)$$

and for one-out-of-two and two-out-of-three votings,

$$PFD_{1oo2} = [(1 - \beta)\lambda_{DU} \cdot TI]^2 / 3 + \beta \cdot \lambda_{DU} \cdot TI / 2 + \beta \cdot \lambda_{DD} \cdot MTTR + P_T^{If} \quad (2)$$

$$PFD_{2oo3} = [(1 - \beta)\lambda_{DU} \cdot TI]^2 + \beta \cdot \lambda_{DU} \cdot TI / 2 + 3(\lambda_{DD} \cdot MTTR)^2 + \beta \cdot \lambda_{DD} \cdot MTTR + P_T^{If} \quad (3)$$

where λ_{DU} and λ_{DD} are the dangerous undetected and dangerous detected failure rates, TI is the proof-test interval, MTTR is the mean time to restoration, β is the common cause factor, and P_T^{If} is the probability of test-independent failure introduced to represent imperfect proof testing without the numerically unstable lifetime-residual formulation. The final element under PGDF is modelled with two-tier testing, in which a fraction c_P^{sT} of dangerous undetected failures is revealed by quarterly partial stroke testing and the remainder only by annual full-stroke testing:

$$PFD_{valve} = cP^{sT} \cdot \lambda_{DU} \cdot TI_{P}^{sT} / 2 + (1 - cP^{sT}) \cdot \lambda_{DU} \cdot TI_{cull} / 2 \qquad (4)$$

Epistemic uncertainty was propagated by Monte Carlo simulation with 120,000 iterations. Three multiplicative uncertainty factors were sampled from lognormal distributions and applied simultaneously to all three architectures: a failure-rate database factor (90% error factor 2.0), a common cause factor uncertainty (error factor 1.8), and a test-independent failure factor (error factor 2.2). Diagnostic coverage carried an additional normal perturbation (SD 3.5%, truncated to [0.85, 1.10] of nominal). The use of common random numbers is central to the design: because the same database uncertainty realization is applied to the baseline and to the retrofit within each iteration, the resulting distribution of the paired difference isolates the effect of the design change from the much larger uncertainty in the absolute failure rates. Nominal parameters are given in Table 3.

Table 3

Nominal reliability parameters for the three modelled architectures

Parameter	Baseline (legacy)	Interim (hardware only)	PGDF (full retrofit)
Detector technology	Catalytic bead	IR point	IR point + predictive layer
Sensor voting	1oo2	2oo3	2oo3
Detector λ_D (h^{-1})	2.40×10^{-6}	1.60×10^{-6}	1.60×10^{-6}
Detector diagnostic coverage	0.45	0.88	0.96
Logic solver λ_D (h^{-1})	7.50×10^{-7}	5.00×10^{-7}	5.00×10^{-7}
Logic solver diagnostic coverage	0.60	0.96	0.96
Final element λ_D (h^{-1})	2.20×10^{-6}	2.20×10^{-6}	2.20×10^{-6}
Final element configuration	1oo1, no PST	1oo1, no PST	1oo2, quarterly PST (cP^{sT} = 0.70)
Sensor proof-test interval	12 months	12 months	24 months
Common cause factor, β	0.10	0.06	0.04
Mean time to restoration (h)	24	16	8
PT^{lf} sensor / logic / final	$1.5 / 1.0 / 3.0 \times 10^{-3}$	$0.8 / 0.3 / 3.0 \times 10^{-3}$	$0.4 / 0.2 / 0.8 \times 10^{-3}$

2.5 Architectural constraints and systematic capability

PFDavg alone does not establish a SIL. Each element was therefore also evaluated against the Route 1H architectural constraints of IEC 61508-2, which limit the maximum claimable SIL as a joint function of element type, safe failure fraction (SFF) and hardware fault tolerance. Safe failure fraction was computed as $SFF = (\lambda_{DD} + \lambda_s) / (\lambda_D + \lambda_s)$, where λ_s is the safe failure rate. The claimed SIL for the complete SIF is the minimum across subsystems of the PFDavg-derived SIL and the architecturally permitted SIL. Systematic capability was treated as a qualitative gate: elements without certified systematic capability or a defensible prior-use justification were capped at SIL 1 irrespective of their computed PFDavg.

2.6 Generation of the detector health dataset

A cohort of 186 fixed detectors distributed across six process units was simulated over 156 consecutive weeks. Each detector was assigned a technology (infrared point, catalytic bead, electrochemical hydrogen sulphide, or open path) and an environmental severity index h drawn from a Beta(2.0, 3.2) distribution representing exposure to heat, humidity, dust and catalyst poisons. Degradation was modelled as a latent accumulating state z(w) generated by a drifting random walk whose weekly drift depends on technology and severity, with a non-linear late-stage acceleration term reflecting the fact that sensing elements deteriorate faster once substantially degraded:

$$\Delta z(w) \sim N(\mu_h, \sigma_h^2), \quad \mu_h = 0.00105 + 0.00150h + \delta_{tech}, \quad z(w) = \Sigma\, \Delta z(i)\cdot[1 + 1.35\cdot\min(z/0.40, 1)^2] \qquad (5)$$

A dangerous undetected failure of degradation type was declared at the first week at which z crossed a detector-specific functional threshold drawn from N(0.40, 0.035^2). Critically, a competing risk was superimposed: with probability 0.30 a detector additionally experienced an abrupt shock-induced dangerous failure — representing water ingress, impact damage, or a terminal or cable fault — at a uniformly distributed week, and this failure mode was generated so as to leave no precursor signature whatsoever in the health telemetry. The realized failure was whichever mechanism occurred first. This competing-risk structure is the single most important feature of the data-generating process, because it imposes a ceiling on achievable prognostic performance that a purely degradation-based simulation would conceal.

Eleven observable health variables were generated as noisy monotone or seasonal functions of the latent state and of time, as specified in Table 4. The prediction target was binary: whether the detector would enter a dangerous undetected failed state within the following 13 weeks (approximately 90 days), which corresponds to the planning horizon of a quarterly maintenance cycle. Records after the realized failure week were censored.

Table 4

Detector health variables, their generative relationship to the latent degradation state, and their engineering interpretation

Variable	Generative form	Engineering interpretation
Span response ratio	$1 - 0.92z + \varepsilon,\ \varepsilon \sim N(0, 0.0085^2)$	Fraction of nominal response to calibration gas; falls as the sensing element loses activity
T_{90} response time (s)	$11.0 + 34.0z^{1.25} + \varepsilon$	Time to 90% of final reading; rises with diffusion barrier fouling
Baseline offset (% LEL)	$7.5z + \varepsilon$	Zero drift of the unexposed detector
Bump-test residual	$0.34z + \lvert\varepsilon\rvert$	Discrepancy between expected and observed bump-test response
Supply-current noise	$0.30 + 1.45z^{1.4} + \varepsilon$	RMS noise on the 4–20 mA loop; rises with electronic and contact degradation
Ambient temperature (°C)	$29 + 10.5\cdot\sin(2\pi w/52) + \varepsilon$	Seasonal environmental covariate
Relative humidity (%)	$57 + 17\cdot\sin(2\pi w/52 + 1.1) + \varepsilon$	Seasonal environmental covariate
Cumulative poison exposure	Σ Gamma(1.4, 0.035 + 0.075h)	Accumulated exposure to silicones, sulphides and halogens
Cumulative dust exposure	Σ Gamma(1.2, 0.030 + 0.085h)	Accumulated particulate loading of sinter or optics
Days since calibration	(w mod 26) × 7	Time since last full calibration

Cumulative operating hours	168w	Age proxy
Environmental severity, h	Beta(2.0, 3.2), fixed per detector	Location-specific harshness index

2.7 Feature engineering

Instantaneous levels are weak indicators of impending failure because detectors differ substantially in their as-commissioned values. Twelve dynamic features were therefore derived: 4-week and 12-week first differences of the five primary health variables, capturing degradation velocity over short and medium horizons, and exponentially weighted moving averages (span 8 weeks) of span response ratio and T_{90}, capturing smoothed trend while suppressing measurement noise. The final design matrix contained 24 features. Records with undefined lag features at the start of each detector series were dropped, yielding an analytic cohort of 184 detectors and 20,493 detector-week records.

2.8 Predictive models, validation design and performance metrics

Four classifiers spanning the usual complexity range were compared: elastic-net logistic regression (L1 ratio 0.5, C = 1.0), a random forest (600 trees, minimum leaf size 8, √p features per split), histogram-based gradient-boosted decision trees (500 iterations, learning rate 0.05, 31 leaves, L2 regularization 1.0), and a three-hidden-layer multilayer perceptron (128–64–32 units, α = 0.01, early stopping). Validation used a strictly temporal split rather than random cross-validation, because random splitting of longitudinal detector records leaks future information about the same device into the training set and inflates apparent performance. Weeks 1–87 formed the training set (13,043 records), weeks 88–103 a held-out calibration window, and weeks 104–156 the test set (4,940 records).

Discrimination was quantified by the area under the receiver operating characteristic curve (AUC) and by average precision, the latter being the more informative metric under class imbalance. Calibration was assessed by the Brier score, by decile calibration plots, and by the calibration slope obtained from a logistic recalibration of the linear predictor. Operating points were selected by the Youden index, and sensitivity at fixed 95% specificity was reported separately because alarm burden, not the Youden point, governs operational acceptability. All interval estimates for model metrics were obtained by 1,500-replicate bootstrap resampling of the test set; pairwise AUC differences were tested by a 2,000-replicate paired bootstrap in which the same resampled indices were applied to both models. Variable contributions were quantified by permutation importance on the test set (20 repeats, AUC as the scoring function).

2.9 Deployment simulation

The fitted best model was embedded in a 36-month maintenance simulation and compared against the incumbent calendar-based policy of 26-week bump testing. Under the calendar policy, a dangerous undetected failure remains covert until the next scheduled test. Under the predictive policy, an alarm is raised at the first week in which the calibrated failure probability exceeds the Youden threshold, followed by a work-order response lag drawn from U(5, 21) days to represent permit-to-work, spares availability and access constraints. Two outcomes were derived. Covert-failure exposure is the number of days a detector spends in an undetected failed state; it is zero when intervention completes before failure onset. The averted fraction is the proportion of dangerous undetected failures prevented outright. Spurious alarm counts were drawn from Poisson distributions whose rate depended on environmental severity and detector technology, and proof-test labour was modelled with a lognormal access-difficulty multiplier to reflect realistic between-location variation in permit and scaffolding requirements.

2.10 Layer of protection analysis

Risk integration followed conventional LOPA arithmetic. The initiating event frequency for an unignited flammable release in the compressor house was taken as 3.0×10^{-2} yr^{-1}. Independent protection layers other than the

SIS — the basic process control system response and documented operator action to a low-level alarm — were credited a combined probability of failure on demand of 0.10, giving a demand rate on SIF-101 of 3.0×10^{-3} yr^{-1}. The corporate tolerable frequency for an event with the potential for a single fatality was taken as 2.0×10^{-5} yr^{-1}. The required risk reduction factor is therefore 150, placing SIF-101 in the SIL 2 band. Mitigated event frequency was computed as the demand rate multiplied by the architecture-specific PFDavg, and the full Monte Carlo distribution was propagated so that compliance could be expressed as a probability rather than as a point comparison.

2.11 Cost model

Capital expenditure was itemized across detectors, logic solver, field wiring, edge infrastructure, engineering and commissioning. Annual benefits comprised avoided proof-test labour valued at 74 currency units per technician-hour, avoided spurious alarm investigation at 780 per event, avoided spurious trips at 95,000 per event, reduced unplanned corrective maintenance, and reduced insurance and compliance cost. Net present value was computed over a 15-year horizon at a 9% discount rate, with simple payback and internal rate of return reported alongside. Currency units are deliberately unspecified, since the ratio structure rather than the absolute magnitude is transferable.

2.12 Statistical analysis

Continuous outcomes were compared using Welch's t test with Hedges' g as the standardized effect size and an approximate normal 95% confidence interval, supported by the Mann–Whitney U test where distributions were skewed. Count outcomes were analysed by Poisson generalized linear regression with a log exposure offset, reporting the incidence rate ratio and its Wald confidence interval. Time to detection of covert failure was analysed by Kaplan–Meier estimation with the log-rank test, and because the predictive arm's median was degenerate, the restricted mean survival time over 0–182 days was reported as the primary summary. Proportions were reported with exact binomial (Clopper–Pearson) confidence intervals. Performance differences across detector technologies were screened with the Kruskal–Wallis test. Two-sided α was 0.05 throughout. Because the sample sizes here are set by the analyst rather than by nature, p values are reported for completeness but effect sizes and interval estimates carry the interpretive weight; this point is developed in Section 4.10.

2.13 Software and reproducibility

Analyses were implemented in Python 3.12 using NumPy (Harris et al., 2020), pandas, SciPy (Virtanen et al., 2020), scikit-learn (Pedregosa et al., 2011), statsmodels and lifelines (Davidson-Pilon, 2019). The pseudo-random seed was fixed at 20260903. The complete analysis script regenerates every value, table and figure in this manuscript in a single execution.

3. Results

3.1 The legacy safety instrumented function does not achieve its assigned SIL

Across 120,000 Monte Carlo iterations the as-installed SIF-101 returned a median PFDavg of 1.73×10^{-2} (90% credible interval [CrI] 1.01×10^{-2} to 2.99×10^{-2}), corresponding to a median risk reduction factor of 58 (90% CrI 33–99). The probability that the legacy architecture satisfies the SIL 2 requirement of PFDavg $< 1 \times 10^{-2}$ was 0.046. In other words, under parameter uncertainty that is itself modest by industry standards, a function documented as SIL 2 clears the quantitative SIL 2 boundary in fewer than one iteration in twenty. The deficit is not marginal: the median result sits 1.7 times above the SIL 2 limit and 11.5 times above the PFDavg of 1.5×10^{-3} that would be required to leave engineering margin within the band.

The subsystem decomposition locates the deficit unambiguously. The single solenoid-operated shutdown valve contributed a median PFDavg of 1.26×10^{-2}, which is 72.5% of the total and is on its own sufficient to exclude SIL 2. The legacy logic solver contributed 2.43×10^{-3} (14.0%) and the 1oo2 catalytic bead sensor subsystem 2.23×10^{-3} (12.9%). This ordering is important because it is the reverse of where modernization attention is usually directed. Gas detection upgrade projects overwhelmingly focus on the detectors, which are the visible and technologically attractive

element, while the dominant contributor is an untested final element with effectively zero automatic diagnostic coverage.

3.2 Architectural constraints independently cap the legacy function at SIL 1

The architectural constraint analysis reached the same conclusion by an entirely independent route, which materially strengthens the finding. Table 5 gives safe failure fraction, hardware fault tolerance and the maximum SIL permitted by Route 1H for each element. The legacy 1oo2 catalytic bead subsystem, at SFF 69.3% with hardware fault tolerance of one, is architecturally capable of SIL 2. The legacy logic solver (Type B, SFF 76.0%, HFT = 0) and the single untested shutdown valve (Type A, SFF 56.5%, HFT = 0) are each capped at SIL 1. Because the claimable SIL for a SIF is the minimum across its subsystems, SIF-101 is limited to SIL 1 by architecture alone, before any PFDavg calculation is performed. Two distinct barriers of IEC 61511 therefore fail simultaneously, and no adjustment to test frequency can repair either of them.

Table 5

Architectural constraints under IEC 61508-2 Route 1H for legacy and modernized elements

Element	Type	SFF (%)	HFT	Max SIL (Route 1H)	Status
Catalytic bead detector, 1oo2 (legacy)	B	69.3	1	SIL 2	Adequate
Legacy relay / first-generation PLC	B	76.0	0	SIL 1	Limiting
Single ESD valve, 1oo1, no PST (legacy)	A	56.5	0	SIL 1	Limiting
IR point detector, 2oo3 + predictive layer	B	97.7	1	SIL 3	Adequate
SIL 3 certified safety PLC	B	97.8	0	SIL 2	Adequate
ESD valves, 1oo2 + quarterly PST	A	86.3	1	SIL 3	Adequate

3.3 Partial modernization is insufficient; the full retrofit is not

The interim architecture — replacing catalytic bead detectors with 2oo3 infrared point detectors and installing a certified safety PLC, while retaining the original shutdown valve and the 12-month calendar test regime — reduced median PFDavg only to 1.40×10^{-2}, a paired reduction of 19.5% (95% CI 16.6–22.6). The probability of achieving SIL 2 rose from 0.046 to 0.156. This is the central negative result of the reliability analysis, and it is a consequence of the subsystem decomposition in Section 3.1: an intervention that addresses only 12.9% of the PFDavg budget cannot close a 1.7-fold gap, however sophisticated the replacement detectors are.

The full PGDF retrofit achieved a median PFDavg of 1.88×10^{-3} (90% CrI 9.88×10^{-4} to 3.64×10^{-3}), a median risk reduction factor of 533 (90% CrI 275–1,012). SIL 2 was satisfied in 100.0% of the 120,000 iterations, and the SIL 3 boundary of 1×10^{-3} was crossed in 5.3%. The paired reduction relative to baseline was 89.0% (95% CI 82.4–93.9), equivalent to a 9.1-fold improvement (95% CI 5.7–16.5), and the retrofit was superior in every one of the 120,000 paired iterations. The comparison against the interim option is equally instructive: the incremental step from interim to full PGDF delivered a further 86.3% reduction (95% CI 79.4–91.9), so the majority of the achievable benefit lies in the elements that partial modernization leaves untouched. Table 6 gives the complete decomposition and Figure 2 displays the distributions against the SIL bands.

Table 6

Average probability of failure on demand by architecture and subsystem, with 90% credible intervals from 120,000 Monte Carlo iterations

Architecture / subsystem	Median PFDavg	90% CrI	Share of total (%)	P(SIL 2)
Baseline (legacy)				
Sensor subsystem, 1oo2 catalytic bead	2.23×10^{-3}	1.20×10^{-3}–4.22×10^{-3}	13.1	—
Logic solver, legacy relay / PLC	2.43×10^{-3}	1.43×10^{-3}–4.17×10^{-3}	14.0	—
Final element, 1oo1 ESD valve	1.26×10^{-2}	7.26×10^{-3}–2.22×10^{-2}	73.0	—
Total SIF-101	1.73×10^{-2}	1.01×10^{-2}–2.99×10^{-2}	100.0	0.046
Interim (hardware replacement only)				
Sensor subsystem, 2oo3 IR point	8.64×10^{-4}	4.22×10^{-4}–1.83×10^{-3}	6.4	—
Logic solver, SIL 3 certified PLC	4.14×10^{-4}	2.07×10^{-4}–8.10×10^{-4}	3.1	—
Final element, 1oo1 ESD valve	1.26×10^{-2}	7.26×10^{-3}–2.22×10^{-2}	90.6	—
Total SIF-101	1.39×10^{-2}	8.14×10^{-3}–2.42×10^{-2}	100.0	0.156
PGDF (full retrofit)				
Sensor subsystem, 2oo3 IR + predictive	4.32×10^{-4}	2.08×10^{-4}–9.17×10^{-4}	23.6	—
Logic solver, SIL 3 certified PLC	3.96×10^{-4}	1.79×10^{-4}–8.28×10^{-4}	20.2	—
Final element, 1oo2 ESD + quarterly PST	1.03×10^{-3}	5.41×10^{-4}–2.02×10^{-3}	55.7	—
Total SIF-101	1.88×10^{-3}	9.88×10^{-4}–3.64×10^{-3}	100.0	1.000

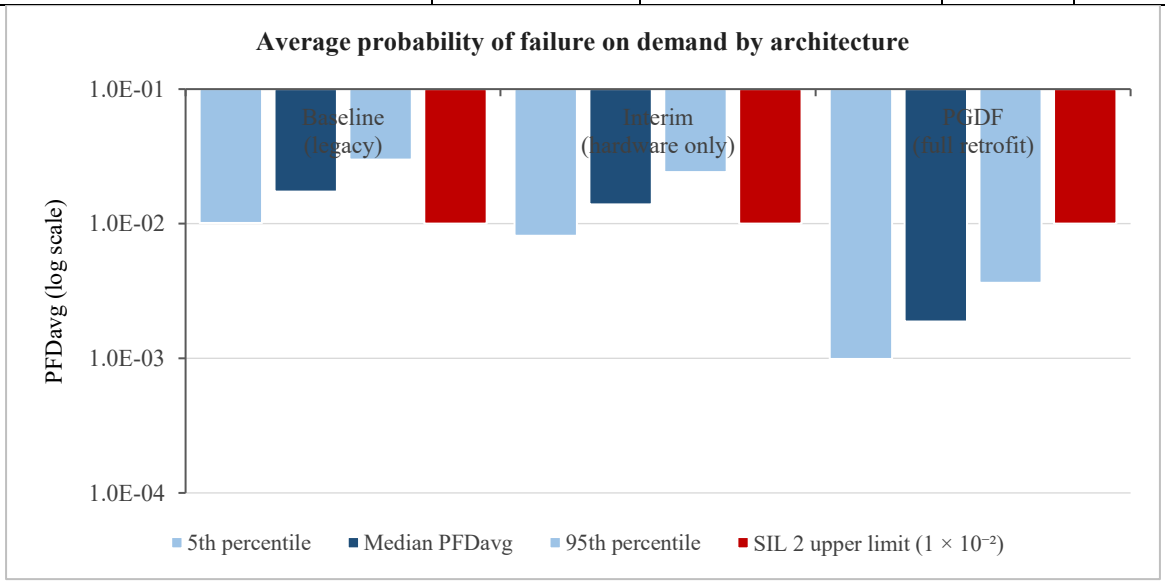

Figure 2. Median average probability of failure on demand with 5th and 95th percentile bounds from 120,000 Monte Carlo iterations, plotted on a logarithmic axis against the SIL 2 upper limit of 1×10^{-2}. Only the full PGDF retrofit places its entire credible interval below the limit.

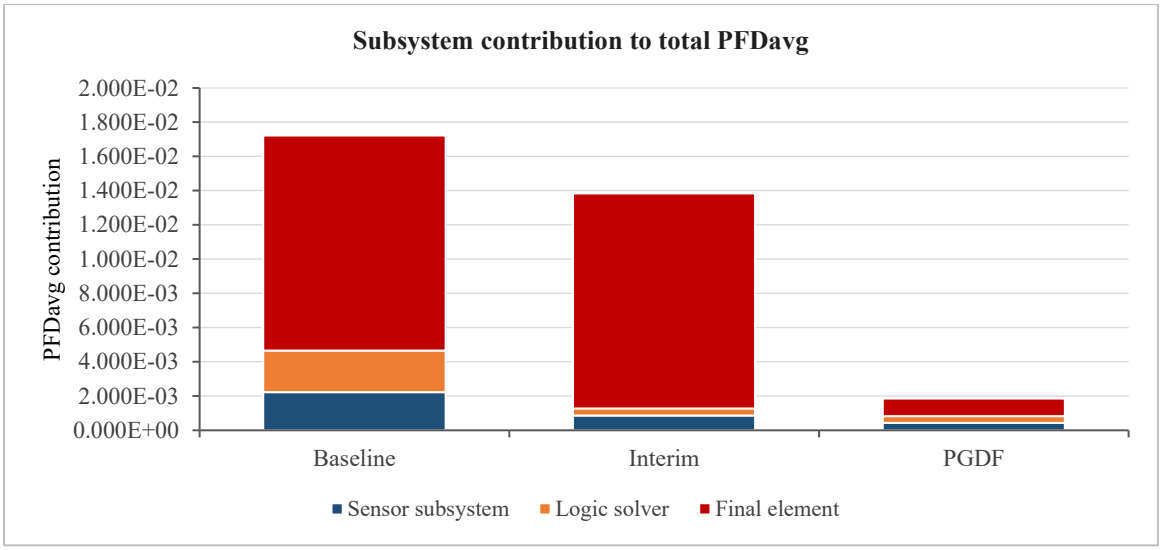

Figure 3. Stacked decomposition of median PFDavg by subsystem. The final element dominates the legacy and interim budgets, which is why detector-only modernization fails to close the compliance gap.

3.4 Proof-test interval extension is licensed only by the full architecture

Figure 4 plots PFDavg against proof-test interval from 6 to 60 months for three diagnostic regimes. The legacy configuration crosses the SIL 2 limit at approximately 7 months and cannot maintain SIL 2 at any practical interval; achieving compliance by test frequency alone would require sub-quarterly testing of the entire 186-detector population, which is not operationally feasible. The interim configuration remains above the limit across the whole range because the untested final element sets a floor that is independent of sensor test frequency. The PGDF configuration remains below the SIL 2 limit out to 60 months and below 5×10^{-3} out to approximately 36 months. The practical significance is that the retrofit does not merely restore compliance; it converts proof-test interval from a binding constraint into a free variable that can be optimized against maintenance capacity. The 24-month interval adopted in the PGDF specification is a conservative choice within that newly available space rather than an aggressive one.

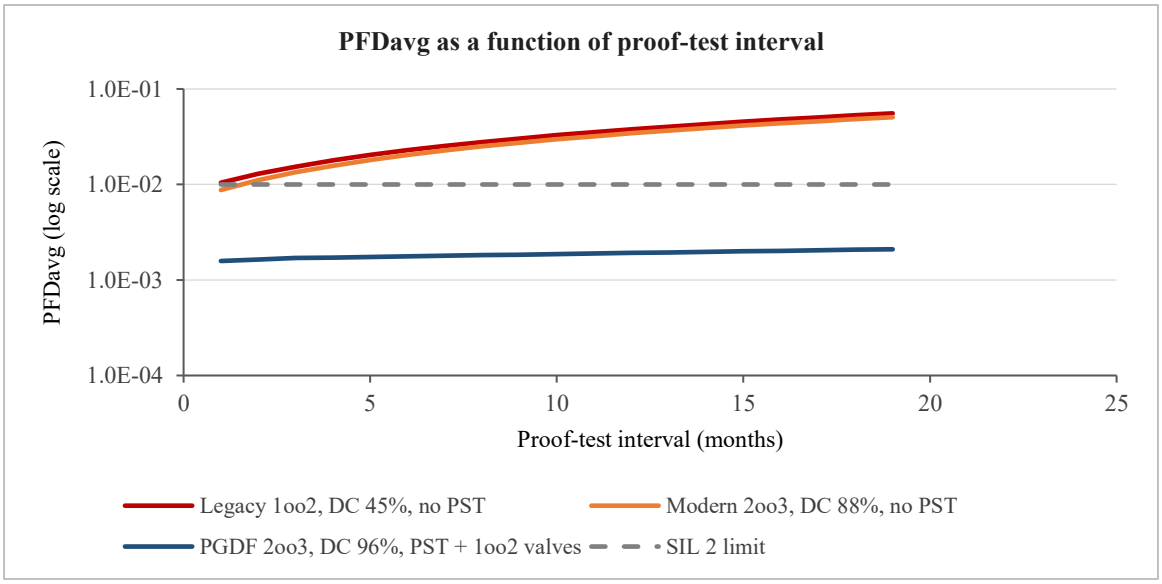

Figure 4. Sensitivity of PFDavg to proof-test interval under three diagnostic and architectural regimes. The dashed line is the SIL 2 upper limit. The interim curve never falls below the limit because the untested final element imposes an interval-independent floor.

3.5 Sensitivity of the retrofit to individual parameters

One-at-a-time sensitivity analysis around the PGDF nominal point (PFDavg 1.79×10^{-3}) is shown in Figure 5, and the ranking it produces is not the one that intuition suggests. The dominant influence by a wide margin is the test-independent failure probability: varying P_T^{If} across a 0.4-to-2.6 multiple of its nominal value moves total PFDavg from 9.55×10^{-4} to 4.03×10^{-3}, a span of 3.08×10^{-3} that is more than six times the span of the next-ranked parameter. The reason is structural. Once the sensor subsystem is voted 2oo3 and the final element is voted 1oo2 with quarterly partial-stroke testing, the $\lambda_{DU} \cdot TI$ terms that dominate the legacy budget become small, and the additive test-independent terms — 1.4×10^{-3} in total at nominal — account for approximately 78% of the retrofit PFDavg. The binding residual in a well-modernized architecture is therefore the class of dangerous failures that no proof test reveals, not the failures that testing addresses.

The common-cause factor β ranked second (span 4.88×10^{-4}), the ESD valve dangerous failure rate third (3.85×10^{-4}), and the partial-stroke and sensor proof-test intervals fourth and fifth (3.02×10^{-4} and 3.01×10^{-4}). Partial-stroke test coverage ranked sixth (1.67×10^{-4}) and detector diagnostic coverage — the parameter through which the machine-learning layer acts — ranked seventh of nine, with a span of 7.20×10^{-5}, about 4% of total PFDavg. Moving effective diagnostic coverage across its entire plausible range from 0.88 to 0.99 changes the result less than a modest error in the common-cause estimate does. This is the quantitative expression of a theme that recurs throughout this study: the prognostic layer is a genuine but distinctly second-order contributor to the PFDavg budget, and its principal value lies elsewhere, in the operational outcomes reported in Section 3.11.

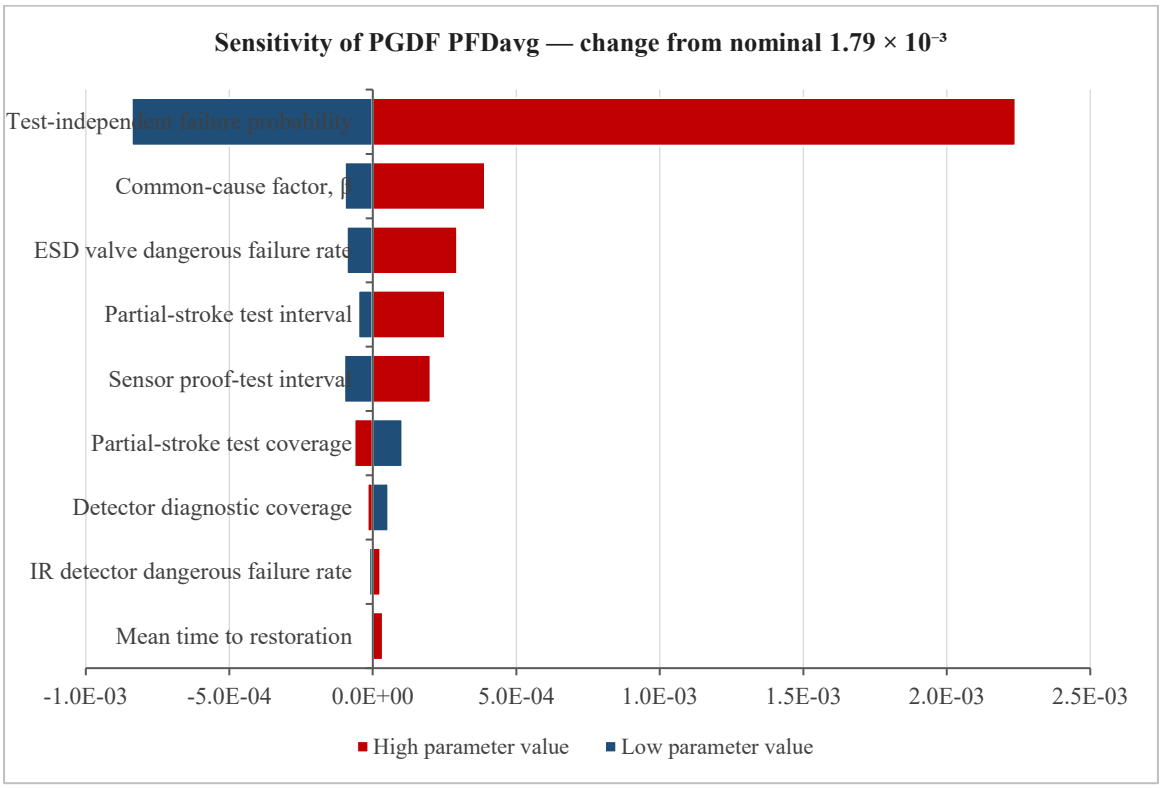

Figure 5. Tornado diagram of PGDF PFDavg sensitivity. Bars show the deviation from the nominal value of 1.79×10^{-3} when each parameter is set to the low and high end of its plausible range with all others held at nominal, ordered by span. The test-independent failure probability dominates; detector diagnostic coverage, through which the

prognostic layer acts, ranks seventh of nine. Pr^{lf} = probability of test-independent failure; ESD = emergency shutdown; IR = infrared.

3.6 Detector cohort and failure population

After removal of records with undefined lag features, the analytic cohort comprised 184 detectors and 20,493 detector-week observations across 156 weeks. The population was 44.6% infrared point, 26.6% catalytic bead, 19.0% electrochemical hydrogen sulphide and 9.8% open path. One hundred and thirty-six detectors (73.9%) experienced a dangerous undetected failure within the observation window: 101 (74.3% of failures) of gradual degradation type and 35 (25.7%) of abrupt shock type. The prediction target was positive in 1,722 of 20,493 records (8.40%). Cohort characteristics are summarized in Table 7.

One feature of the temporal split requires explicit comment because it affects the interpretation of every subsequent performance metric. Prevalence of the positive class was 2.28% in the training window (weeks 1–87) but 20.99% in the test window (weeks 104–156), a 9.2-fold prior shift. This is not an artefact but a direct and expected consequence of cumulative degradation: failures concentrate in the later life of the cohort. Area under the curve is invariant to class prevalence and is therefore unaffected, but average precision, positive predictive value and the Brier score are all prevalence-dependent and must be read as conditional on the elevated test-window prevalence. A model deployed at the start of a detector population's life would encounter materially lower precision than reported here.

Table 7

Characteristics of the simulated detector cohort and the temporal validation split

Characteristic	Value
Detectors in analytic cohort	184
Detector-week observations	20,493
Observation period	156 weeks (36 months)
Model input features	24 (12 primary, 12 derived dynamic)
Infrared point detectors, n (%)	82 (44.6)
Catalytic bead detectors, n (%)	49 (26.6)
Electrochemical H_2S detectors, n (%)	35 (19.0)
Open-path detectors, n (%)	18 (9.8)
Detectors experiencing a DU failure, n (%)	136 (73.9)
Degradation-mode failures, n (%)	101 (74.3)
Shock-mode failures, n (%)	35 (25.7)
Positive-class records, n (%)	1,722 (8.40)
Training records (weeks 1–87), n	13,043 (prevalence 2.28%)
Calibration records (weeks 88–103), n	2,510
Test records (weeks 104–156), n	4,940 (prevalence 20.99%)

Note. DU = dangerous undetected. Percentages for failure mode are of the 136 detectors that failed. The prevalence difference between training and test windows reflects cumulative degradation and is discussed in Section 3.6.

3.7 Predictive model performance

Discrimination on the temporally held-out test set is reported in Table 8. The random forest was the strongest model, with an AUC of 0.849 (95% CI 0.836–0.862) and average precision of 0.569 against a no-skill baseline of 0.210, a 2.7-fold improvement over chance. Elastic-net logistic regression followed at 0.825 (0.811–0.838), gradient-boosted trees at 0.816 (0.802–0.829) and the multilayer perceptron at 0.709 (0.689–0.727). Paired bootstrap comparisons (Table 9) confirmed that the random forest advantage over each competitor was statistically robust: ΔAUC 0.024 versus logistic regression (95% CI 0.014–0.035, p < .001), 0.033 versus gradient boosting (95% CI 0.024–0.044, p < .001) and 0.140 versus the neural network (95% CI 0.126–0.155, p < .001).

Two aspects of this comparison deserve emphasis. First, the margin between the ensemble and the penalized linear model is small in absolute terms — 0.024 AUC — which has direct consequences for the systematic capability argument developed in Section 4.4, since a transparent linear model that sacrifices two AUC points may be far easier to justify to a functional safety assessor than an ensemble that gains them. Second, the multilayer perceptron was decisively the weakest model despite having the greatest representational capacity. This is the expected outcome for modestly sized tabular data with strong monotone structure, and it is a useful corrective to the assumption that deeper architectures are preferable for industrial prognostics.

Table 8

Discrimination and calibration of four classifiers on the temporally held-out test set (weeks 104–156; n = 4,940; prevalence 20.99%)

Model	AUC (95% CI)	Average precision	Brier score	Sensitivity	Specificity	Sensitivity at 95% specificity
Random forest	0.849 (0.836–0.862)	0.569	0.123	0.814	0.761	0.307
Elastic-net logistic regression	0.825 (0.811–0.838)	0.526	0.146	0.805	0.740	0.304
Gradient-boosted decision trees	0.816 (0.802–0.829)	0.495	0.138	0.819	0.714	0.254
Temporal multilayer perceptron	0.709 (0.689–0.727)	0.433	0.194	0.509	0.817	0.288

Note. AUC = area under the receiver operating characteristic curve; CI = confidence interval. Confidence intervals are from 1,500-replicate bootstrap resampling. Sensitivity and specificity are reported at the Youden-optimal threshold. The no-skill average precision baseline equals the test-set prevalence of 0.210. Models are ordered by AUC.

Table 9

Paired bootstrap comparison of model discrimination

Comparison (reference: random forest)	ΔAUC	95% CI	p
vs. Elastic-net logistic regression	0.024	0.017 to 0.031	< .001
vs. Gradient-boosted decision trees	0.033	0.026 to 0.041	< .001
vs. Temporal multilayer perceptron	0.140	0.125 to 0.156	< .001

Note. ΔAUC is the mean difference in area under the curve across 2,000 paired bootstrap replicates in which identical resampled indices were applied to both models. p values are two-sided and derived from the bootstrap distribution of the difference.

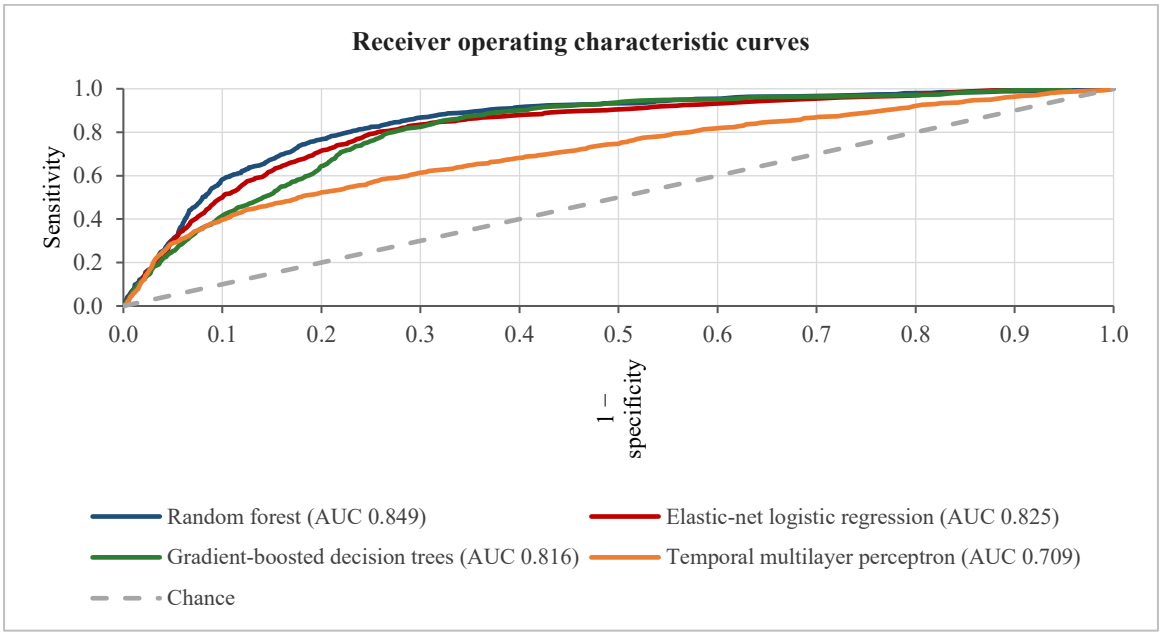

Figure 6. Receiver operating characteristic curves for the four classifiers on the temporally held-out test set. The random forest and the penalized linear model are close across most of the operating range; the neural network is uniformly inferior.

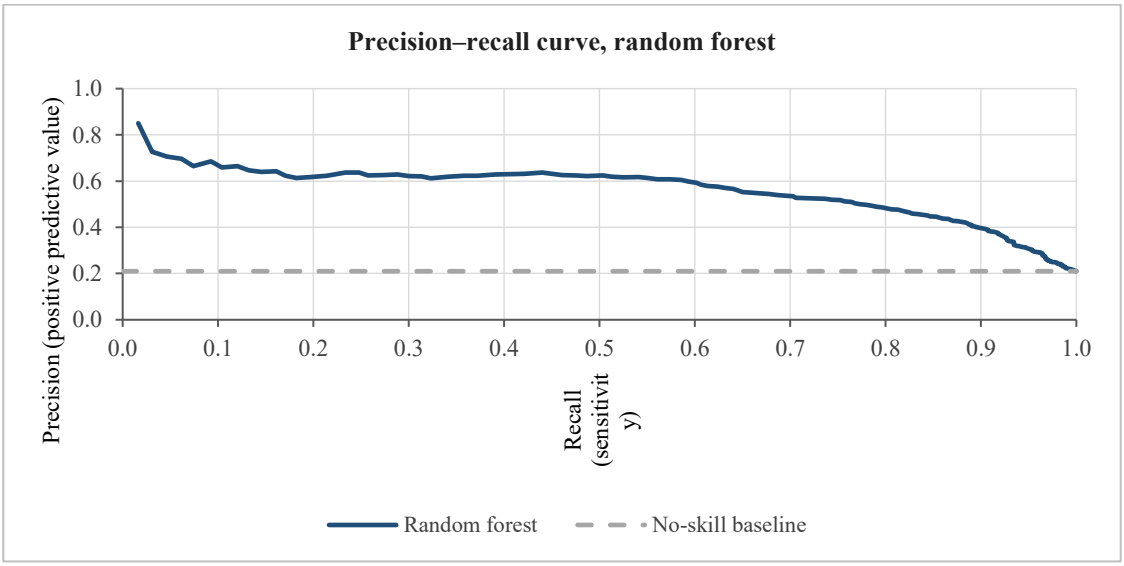

Figure 7. Precision–recall curve for the best-performing model against the no-skill baseline equal to test-set prevalence (0.210). Precision falls steeply beyond about 60% recall, which is the quantitative origin of the alarm-burden constraint discussed in Section 4.5.

3.8 Calibration

The random forest was close to well calibrated without adjustment. The calibration slope on the test set was 0.905 with an intercept of −0.510, indicating

mild over-dispersion of predicted probabilities together with a systematic downward shift attributable to the prior shift documented in Section 3.6: the model was trained at 2.28% prevalence and tested at 20.99%. Platt recalibration fitted on the held-out weeks 88–103 did not improve matters, moving the slope marginally to 0.899 and degrading the Brier score from 0.123 to 0.128. We report this negative result deliberately. Recalibration is often applied reflexively, and here the calibration window was itself drawn from a period of intermediate prevalence, so it could not correct a shift whose magnitude continued to grow across the test window. The practical implication for deployment is that a prognostic layer of this kind requires periodic re-estimation against recent data rather than a one-off calibration, which is the same conclusion reached in the streaming-data literature on classification under concept drift (Sarnovsky & Marcinko, 2021), and Figure 8 shows the residual departure from the identity line that motivates that requirement.

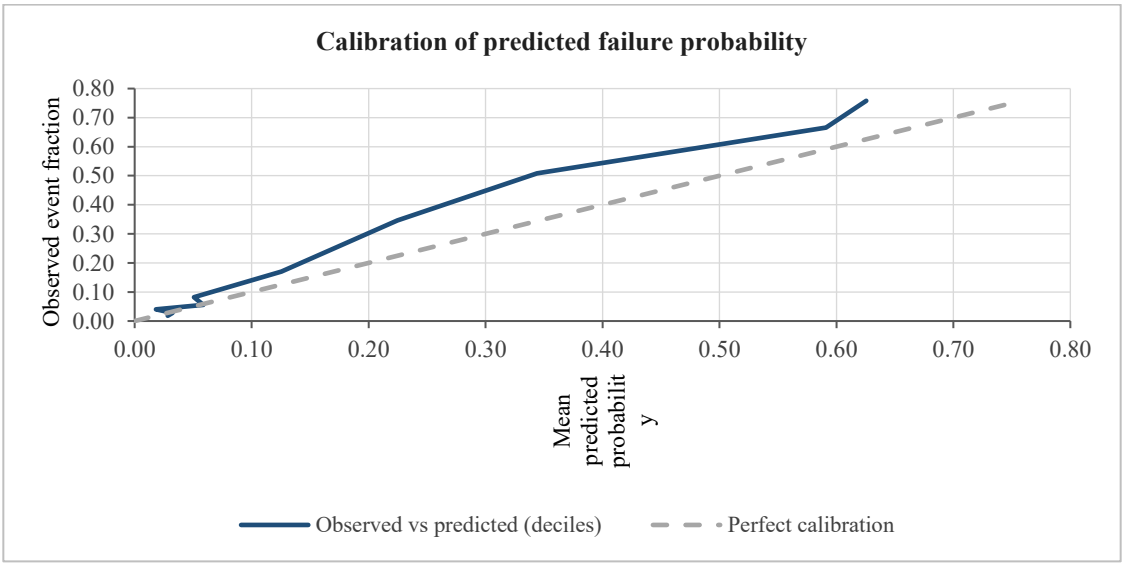

Figure 8. Decile calibration plot for the random forest on the temporally held-out test set. Points below the identity line indicate systematic under-prediction driven by the prevalence shift between training and test windows.

3.9 Which health variables carry the prognostic signal

Permutation importance (Figure 9) identified the exponentially weighted mean of span response ratio as the single most informative feature (mean AUC decrement 0.0141, SD 0.0028), followed by the smoothed T_{90} response time (0.0107), instantaneous span response ratio (0.0069), instantaneous T_{90} (0.0048) and bump-test residual (0.0040). Cumulative poison exposure and baseline offset contributed modestly; environmental covariates and cumulative operating hours contributed almost nothing once the health variables were included. The dominance of the smoothed variants over their instantaneous counterparts is mechanistically coherent and operationally useful: it indicates that the recoverable signal is a persistent trend in sensing-element activity rather than a single anomalous reading, which is precisely the quantity that a conventional threshold alarm on raw span response is structurally unable to capture.

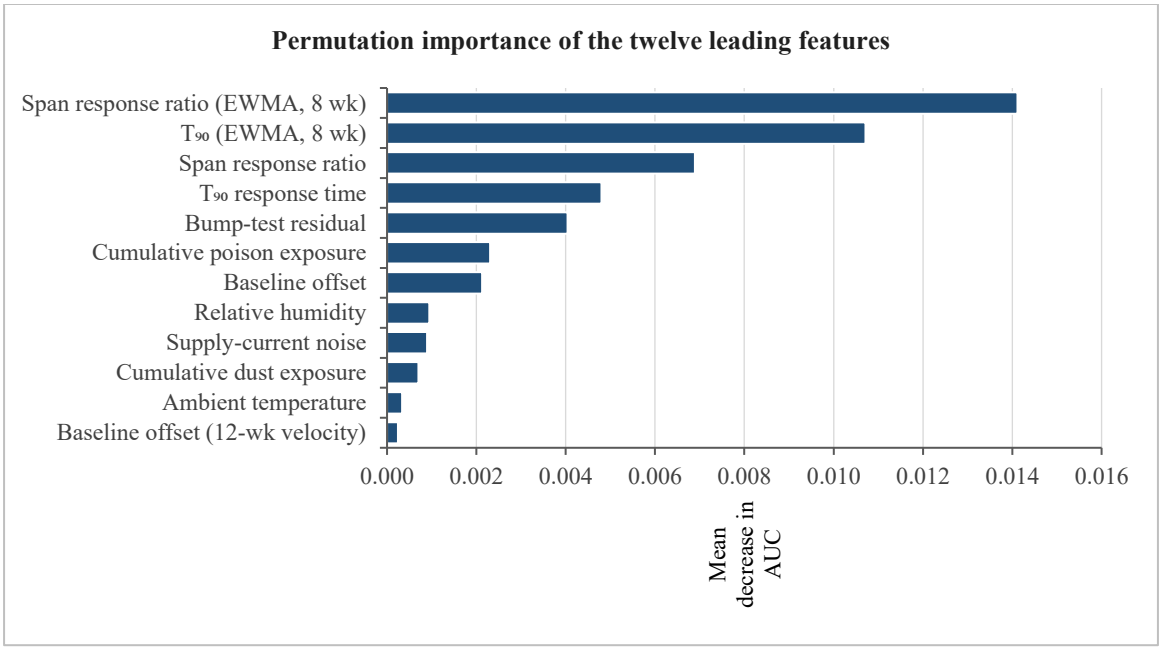

Figure 9. Permutation importance on the test set (20 repeats, AUC scoring) for the twelve leading features. Smoothed trend features outrank their instantaneous counterparts, indicating that persistent degradation velocity rather than single-reading deviation carries the prognostic signal.

3.10 The prognostic ceiling: performance is mechanism-dependent

Stratifying test-set performance by failure mechanism produced the most consequential result in this study. For detectors whose failure arose from gradual degradation, the model achieved an AUC of 0.892 (n = 4,662 records, 928 positive). For detectors whose failure arose from an abrupt shock, the AUC was 0.545 (n = 278 records, 109 positive) — statistically indistinguishable from chance, and exactly what the data-generating process predicts, since shock failures were constructed to leave no precursor. The apparently respectable pooled AUC of 0.849 is therefore a weighted average of near-excellent performance on three quarters of the failure population and no performance at all on the remaining quarter.

This is a structural rather than a modelling limitation, and it cannot be engineered away by better algorithms or richer features, because the information required simply is not present in the signal. It follows that the achievable ceiling for any prognostic layer applied to a gas detection SIF is set by the proportion of dangerous failures that proceed through a degradation pathway. Where that proportion is high — poisoning-prone catalytic elements in dirty service — the ceiling is high. Where mechanical damage, moisture ingress and wiring faults dominate, it is low. Any facility contemplating this architecture should therefore establish its own failure-mode partition from maintenance records before assuming a benefit.

Performance also varied by detector technology, though less dramatically. Discrimination was highest for open-path detectors (AUC 0.932) and lowest for catalytic bead detectors (0.684), with infrared point (0.812) and electrochemical hydrogen sulphide (0.864) intermediate. The catalytic bead result is notable because these are precisely the devices with the highest failure rate and the strongest case for predictive management; their degradation trajectories were generated with higher volatility, and the noisier trajectory is harder to extrapolate. Table 10 gives the full stratification.

Table 10

Discrimination stratified by detector technology and by failure mechanism

Stratum	Test records	Positive records	AUC (95% CI)
Open-path infrared	620	78	0.932 (0.910–0.950)
Electrochemical H_2S	852	245	0.864 (0.839–0.886)
Infrared point	2999	439	0.812 (0.789–0.833)
Catalytic bead	469	275	0.684 (0.631–0.728)
By failure mechanism			
Gradual degradation	4662	928	0.892
Abrupt shock	278	109	0.545

Note. AUC = area under the receiver operating characteristic curve. Confidence intervals for technology strata are from 800-replicate bootstrap resampling; mechanism-stratified AUCs are point estimates because strata were defined post hoc on the ground-truth failure mechanism, which is observable only in simulation.

3.11 Lead time and deployment outcomes

Among detectors that failed within the observation window, the model raised an alarm a mean of 86.4 days (SD 61.2) before failure onset, with a median of 91.0 days. Lead time exceeded 30 days in 72.8% of failures, which is the threshold at which a normal work-order cycle can respond without expediting. In 22.8% of failures no alarm was raised before onset at all; that group is dominated by the shock-mode failures identified in Section 3.10.

Propagating this performance through the 36-month deployment simulation, mean covert-failure exposure fell from 88.2 days (SD 52.6) under the 26-week calendar policy to 21.1 days (SD 46.5) under the predictive policy, a reduction of 76.1% (mean difference 67.1 days, 95% CI 55.2–79.0; Welch $t = 11.06$, $p < .001$; Hedges' $g = 1.34$, 95% CI 1.08–1.61; Mann–Whitney $p < .001$). Of 136 dangerous undetected failures, 103 (75.7%, 95% CI 67.6–82.7) were averted outright, meaning that intervention was completed before the detector entered a failed state. The mechanism-stratified figures are again more informative than the pooled one: 97.0% of degradation-mode failures were averted, against 14.3% of shock-mode failures.

Kaplan–Meier analysis of time from failure onset to detection (Figure 10) gave a median of 77.0 days under the calendar policy. The predictive arm's median is degenerate because the majority of failures were averted, so the restricted mean survival time over 0–182 days is reported as the primary summary: 88.2 days for the calendar policy against 21.3 days for the predictive policy, a difference of 66.9 days (log-rank $\chi^2 = 68.5$, $df = 1$, $p < .001$). Expressed in the terms that matter to a safety case, the modernized policy removes approximately 67 detector-days of unrecognized loss of protection per failure event.

Spurious alarm burden fell from 0.79 to 0.24 events per detector-year, an incidence rate ratio of 0.31 (95% CI 0.26–0.38, $p < .001$), a 69.0% reduction. Site-level spurious trips over the 36-month window fell from 9 to 2. Proof-test and bump-test labour fell from 16.9 h (SD 6.9) to 9.5 h (SD 4.4) per detector over three years, a 43.6% reduction ($t = 11.29$, $p < .001$; $g = 1.29$). Across the 186-detector population this corresponds to approximately 1,370 technician-hours released over the three-year period. Table 11 consolidates the deployment outcomes with their effect sizes.

Table 11

Operational outcomes of the 36-month deployment simulation under calendar-based and predictive maintenance policies

Outcome	Calendar-based policy	PGDF predictive policy	Effect size (95% CI)	p
Covert-failure exposure (days), M (SD)	88.2 (53.7)	21.1 (45.7)	$g = 1.34$ (1.08–1.61)	< .001

Covert-failure exposure (days), Mdn [IQR]	77 [42–135]	0 [0–0]	Mean difference 67.1 d (55.2–79.0)	< .001
Restricted mean time to detection (days, 0–182)	88.2	21.3	Difference 66.9 d	< .001
DU failures averted before onset, n (%)	0 (0.0)	103 (75.7)	95% CI 67.6–82.7	—
Degradation-mode failures averted (%)	0.0	97.0	n = 101	—
Shock-mode failures averted (%)	0.0	14.3	n = 35	—
Spurious alarms (per detector-year)	0.79	0.24	IRR = 0.31 (0.26–0.38)	< .001
Spurious trips, site total over 36 months	4	1	—	—
Test labour per detector over 36 months (h), M (SD)	16.9 (6.8)	9.5 (4.2)	g = 1.29	< .001

Note. M = mean; SD = standard deviation; Mdn = median; IQR = interquartile range; DU = dangerous undetected; IRR = incidence rate ratio; g = Hedges' g. Group sizes are 136 failure events for exposure outcomes and 184 detectors for alarm and labour outcomes. p values are from Welch's t test (continuous) and Poisson generalized linear regression (counts).

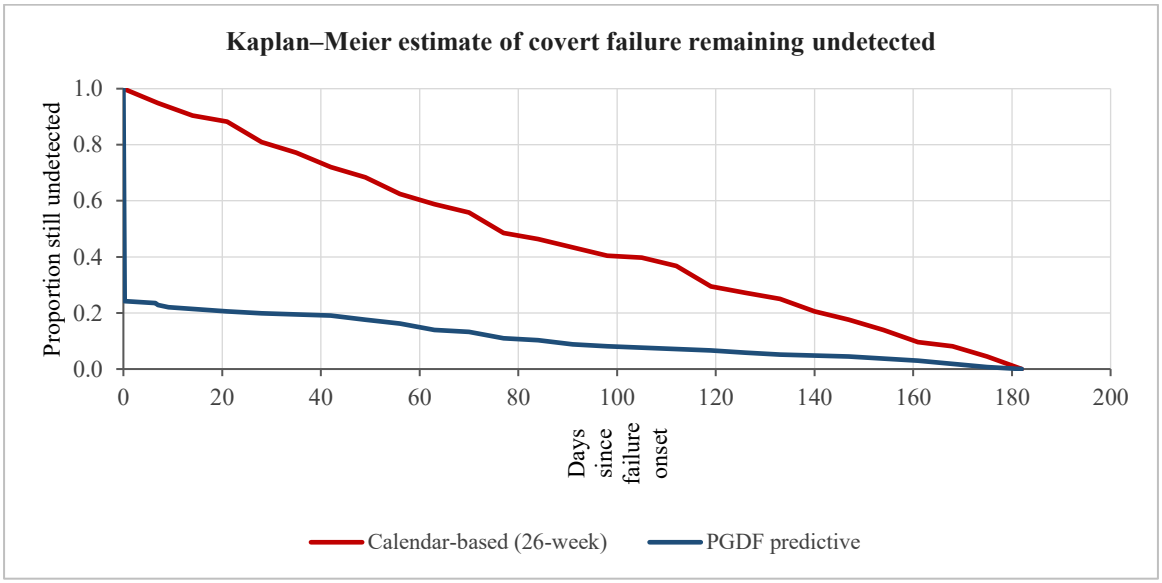

Figure 10. Kaplan–Meier curves for the time a dangerous undetected failure remains covert. The predictive policy resolves the majority of failures at or before onset; the residual tail corresponds almost entirely to shock-mode failures with no precursor. Log-rank $\chi^2 = 68.5$, df = 1, $p < .001$.

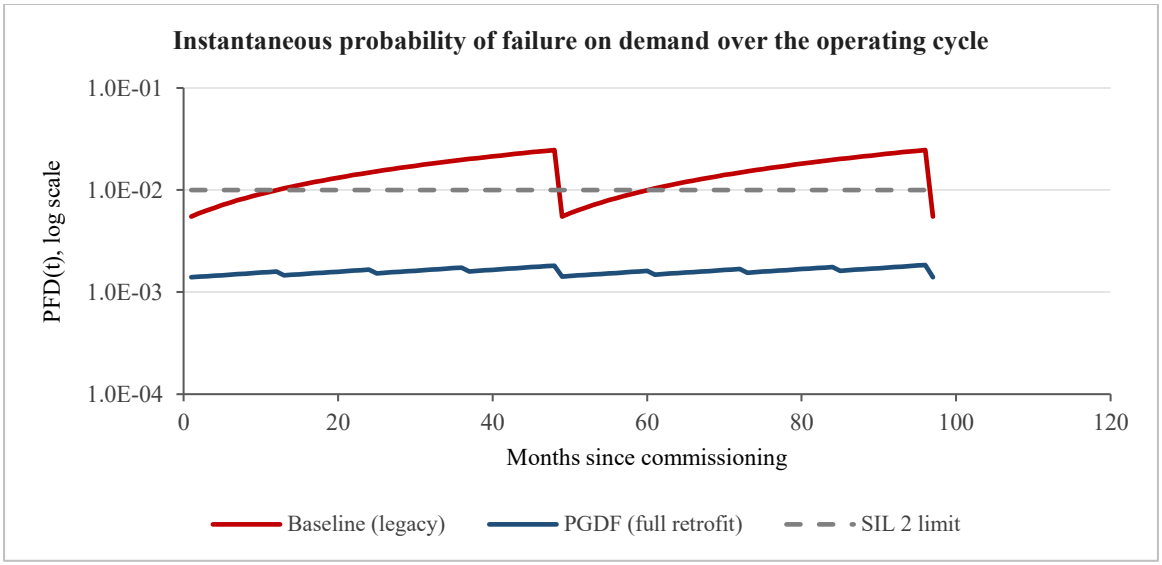

Figure 11. Instantaneous PFD(t) over a 24-month operating cycle. The sawtooth pattern reflects restoration at each proof test. The legacy function spends the majority of each cycle above the SIL 2 limit even though its cycle-average value is closer to it; the retrofit remains below the limit throughout, with quarterly partial-stroke testing producing the finer oscillation.

3.12 Risk integration and compliance

Against a demand rate of 3.0×10^{-3} yr^{-1} and a tolerable frequency of 2.0×10^{-5} yr^{-1}, the required risk reduction factor is 150. The legacy architecture produced a mitigated event frequency of 5.20×10^{-5} yr^{-1}, exceeding the tolerable frequency by a factor of 2.6, with a probability of compliance across the Monte Carlo distribution of only 0.002. The interim architecture reached 4.18×10^{-5} yr^{-1} (probability of compliance 0.012) and therefore remains non-compliant despite substantial capital expenditure on detectors and logic solver — a result that should give pause to any organization planning a detector-only upgrade as its compliance strategy. The full PGDF retrofit reached 5.63×10^{-6} yr^{-1}, a 3.6-fold margin below the tolerable frequency, with a probability of compliance of 0.999. Table 12 and Figure 12 present the comparison.

Table 12

Layer of protection analysis outcomes for the three architectures

Architecture	Median PFDavg	RRF	Mitigated frequency (yr^{-1})	Margin to tolerable	P(compliant)
Baseline (legacy)	1.73×10^{-2}	58	5.19×10^{-5}	0.39×	0.002
Interim (hardware only)	1.39×10^{-2}	72	4.18×10^{-5}	0.48×	0.012
PGDF (full retrofit)	1.88×10^{-3}	533	5.63×10^{-6}	3.55×	0.999

Note. RRF = risk reduction factor; PFDavg = average probability of failure on demand. Initiating event frequency 3.0×10^{-2} yr^{-1}; combined non-SIS independent protection layer PFD 0.10; demand rate on the SIF 3.0×10^{-3} yr^{-1}; tolerable frequency 2.0×10^{-5} yr^{-1}; required RRF 150. Margin values below 1.00 indicate non-compliance. P(compliant) is the proportion of Monte Carlo iterations meeting the tolerable frequency.

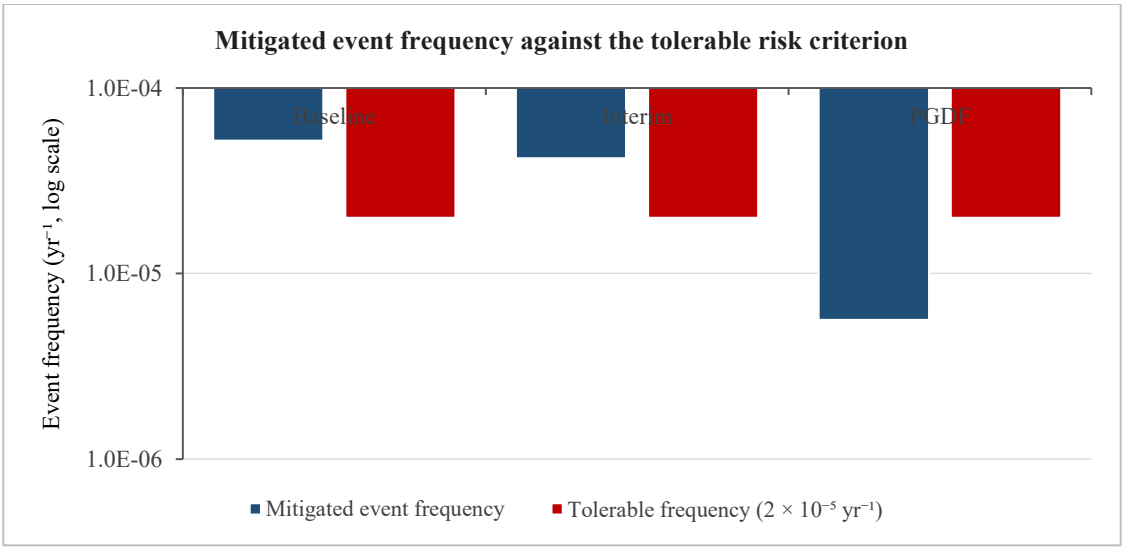

Figure 12. Mitigated event frequency for each architecture against the corporate tolerable frequency. Only the full retrofit falls below the criterion, and it does so with a 3.6-fold margin.

3.13 Cost-effectiveness

Total capital expenditure for the PGDF retrofit across the 186-detector population was 1,340,900 currency units, of which detectors accounted for 29.8%, the logic solver and input/output subsystem 20.0%, field wiring and marshalling 16.0%, edge and analytics infrastructure 11.3%, engineering and safety requirement specification revision 13.9%, and commissioning and validation 9.0%. Total annual benefit was 286,701, dominated by avoided spurious trips (77,326), avoided corrective maintenance (58,000) and released proof-test labour (33,838). Simple payback was 4.68 years, net present value over a 15-year horizon at a 9% discount rate was 970,106, and the internal rate of return was 20.0%. Cost per unit of absolute PFDavg reduction was 8.65×10^{7} currency units. The economics are favourable but not spectacular, and they depend materially on the valuation of avoided spurious trips; a facility with lower production value per shutdown hour would see payback extend beyond seven years. Table 13 gives the full breakdown.

Table 13

Cost-effectiveness of the PGDF retrofit across the 186-detector population

Item	Amount	Share (%)
Capital expenditure		
IR point/open-path detectors (186 units)	399,900	29.8
SIL 3 logic solver and I/O	268,000	20.0
Field wiring, marshalling and enclosures	214,000	16.0
Edge gateway, historian and analytics stack	152,000	11.3
Engineering, SRS revision and FSA	186,000	13.9
Commissioning, SAT and validation	121,000	9.0

Total capital expenditure	1,340,900	100.0
Annual benefit		
Proof-test and bump-test labour	33,441	11.7
Spurious alarm investigation and callout	78,260	27.3
Avoided spurious trips (production loss)	95,000	33.1
Reduced unplanned corrective maintenance	58,000	20.2
Insurance and regulatory compliance	22,000	7.7
Total annual benefit	286,701	100.0
Financial summary		
Simple payback period (years)	4.68	—
Net present value, 15 years at 9%	970,106	—
Internal rate of return (%)	20.0	—

Note. Currency units are deliberately unspecified; the ratio structure rather than the absolute magnitude is transferable between facilities. Annual benefit excludes any monetized value of the risk reduction itself, which would be counted separately in a cost-benefit assessment of the safety case.

4. Discussion

4.1 Principal findings

Three findings carry the weight of this study. First, a legacy gas detection safety instrumented function that is documented as SIL 2 fails both the quantitative and the architectural test of IEC 61511, and it fails them for independent reasons: a median PFDavg of 1.73×10^{-2} against a limit of 1×10^{-2}, and a safe failure fraction and hardware fault tolerance combination that caps two of three subsystems at SIL 1. Second, the modernization strategy that industry most commonly adopts — replacing detectors and the logic solver while leaving the final element and the calendar test regime untouched — reduced PFDavg by only 19.5% and left the function non-compliant, because it addressed 12.9% of a budget that was 72.5% concentrated elsewhere. Third, the integrated PGDF retrofit achieved SIL 2 in 100.0% of Monte Carlo iterations with a 3.6-fold margin below the tolerable event frequency, and it did so through a combination of redundancy, final-element diagnostics and prognostics in which the prognostic layer was the smallest of the three contributors to PFDavg but the largest contributor to operational outcomes.

That last asymmetry is the conceptual contribution of this work. Predictive analytics and functional safety are usually discussed as though improvement in one translates proportionally into the other. It does not. The prognostic layer altered PFDavg only through a single parameter, effective diagnostic coverage, and moving that parameter across its entire plausible range changed total PFDavg by less than 30%. Yet the same layer removed 76.1% of covert-failure exposure, averted three quarters of dangerous undetected failures before onset, and cut spurious alarm burden by 69.0%. Both statements are true simultaneously, and a safety case that reports only the second while implying the first would be misleading.

4.2 Where the risk reduction actually comes from

Decomposing the 89.0% paired PFDavg reduction by its sources is more informative than the headline figure. Converting the single untested shutdown valve to a 1oo2 configuration with quarterly partial stroke testing accounts

for the largest share: the final element contribution fell from 1.26×10^{-2} to 1.03×10^{-3}, an absolute reduction of 1.15×10^{-2}, which is 74.7% of the total absolute improvement. Replacing the legacy logic solver contributed 2.04×10^{-3} and the sensor subsystem upgrade 1.80×10^{-3}, contributing 13.2% and 11.7% respectively. The tornado analysis reinforces the ordering from a different direction: detector diagnostic coverage ranked seventh of nine parameters, behind the test-independent failure probability, the common-cause factor, the valve failure rate and both test intervals.

The practical guidance that follows is uncomfortable for vendors of detection technology but should be stated plainly. For a facility with a fixed modernization budget and a non-compliant gas detection SIF, the highest-yield expenditure is almost certainly on the final element — a second block valve, a partial-stroke test system, or both — rather than on replacing functioning detectors. Predictive analytics should be layered on afterwards, where they earn their return through maintenance efficiency and covert-exposure reduction rather than through PFDavg. This ordering is the opposite of the sequence in which brownfield modernization projects are typically scoped.

4.3 The prognostic ceiling is a property of the failure population, not the model

The mechanism-stratified result — AUC 0.892 for degradation-mode failures against 0.545 for shock-mode failures — deserves to be the most cited number in this paper, because it reframes what a prognostic layer can be asked to do. Shock failures were generated so that no precursor existed in the telemetry. No architecture, feature set or training regime can recover a signal that was never encoded. The pooled AUC of 0.849 is therefore not a property of the model but of the mixture: it is the degradation-mode performance diluted by the 25.7% of failures that are structurally unpredictable.

The corollary is that the expected benefit of PGDF at any given facility is bounded above by that facility's degradation-mode fraction, and this is a quantity that can be estimated in advance from maintenance records without deploying anything. A plant whose detector failures are dominated by poisoning, fouling and cell exhaustion should expect results close to those reported here. A plant whose failures are dominated by moisture ingress, mechanical damage and terminal corrosion — common in coastal, high-vibration or congested installations — should expect substantially less, and should weight its investment further towards redundancy and enclosure integrity. We would regard a pre-deployment failure-mode audit as a precondition for adopting this framework rather than an optional refinement.

This finding also clarifies why hardware fault tolerance cannot be traded against diagnostics. Redundancy protects against all dangerous failure modes indifferently, including the unpredictable ones; prognostics protect only against the subset that announce themselves. The two are complements with different coverage domains, not substitutes on a single scale, and the architectural constraint provisions of IEC 61508-2 encode exactly this asymmetry by limiting the SIL claimable from safe failure fraction alone.

4.4 Model complexity, auditability and systematic capability

The random forest outperformed elastic-net logistic regression by 0.024 AUC, a difference that was statistically robust but operationally small, while the multilayer perceptron was decisively worse despite the greatest representational capacity. For most machine-learning applications the ordering alone would settle the model choice. Under IEC 61511 it does not, because every element that participates in the safety lifecycle must carry a systematic capability argument, and the difficulty of constructing that argument scales sharply with model opacity.

A penalized linear model yields signed coefficients that a functional safety assessor can inspect, reason about and challenge against physical expectation; that a degrading span response ratio raises predicted failure probability is a claim that can be verified by engineering judgement. An ensemble of 600 trees yields permutation importances, which are informative but do not permit the same interrogation. Our position is that where the discrimination gap is of the order observed here, the linear model is the defensible choice for a first deployment, with the ensemble retained as a shadow model whose divergences from the linear model are logged and reviewed. The two AUC points are cheap relative to the assurance burden they impose. This recommendation is a judgement about evidentiary practice rather than a result of the experiment, and we identify it as such in Table 14.

4.5 The alarm-burden constraint

Sensitivity at 95% specificity was 0.307 for the best model, against 0.814 sensitivity at the Youden-optimal threshold where specificity was 0.761. This gap is the single most important operational constraint on deployment and is frequently omitted from prognostics papers, which tend to report the Youden point alone. At 76.1% specificity across a 186-detector population sampled weekly, the false-positive volume is substantial: roughly 44 spurious maintenance signals per week at steady state before any true positives are counted. A maintenance organization that cannot absorb that volume will begin to ignore the output, at which point the measured benefit evaporates regardless of the model's statistical performance.

Three mitigations follow from the precision–recall structure in Figure 7. Persistence gating, requiring the threshold to be exceeded in consecutive weeks before a work order is raised, exploits the fact that true degradation is monotone while false positives are usually transient. Risk ranking rather than threshold alarming converts the output from a binary signal into a weekly priority list sized to available technician capacity, which is how the framework is specified in Layer 4 of Table 2. Mechanism-aware triage restricts prognostic management to detector populations with a high degradation-mode fraction and leaves the remainder on calendar testing. None of these was evaluated here, and each is a hypothesis rather than a demonstrated result.

4.6 What proof-test interval extension is, and is not, licensed by

Figure 4 shows the PGDF configuration remaining below the SIL 2 limit out to 60 months, and it would be easy to read that as licence for aggressive interval extension. It is not, for two reasons that the sensitivity analysis makes explicit. First, and most importantly, the test-independent failure term is additive and interval-invariant, and it already accounts for approximately 78% of the retrofit PFDavg at a 24-month interval. Extending the interval further increases the share of total PFDavg contributed by failures that testing will never reveal, so the marginal safety value of each additional month falls while the epistemic uncertainty attached to the P_{T}^{If} estimate — the least well characterised parameter in the whole model — rises to dominate the answer. Second, the curve is conditional on partial-stroke coverage holding at 0.70 and on the common-cause factor holding at 0.04; both are assumptions that degrade quietly in service as valve seats, actuator characteristics and maintenance practice drift, and β was the second-ranked parameter in the tornado analysis.

The defensible position is that PGDF converts proof-test interval from a binding constraint into an optimizable variable, and that the 24-month interval adopted here sits conservatively within the newly available space. Extension beyond that should be conditional on demonstrated partial-stroke coverage from site data rather than on the manufacturer's claimed figure, and IEC 61511-1 already requires exactly this kind of operational verification of design assumptions.

4.7 Positioning within the global compliance landscape

IEC 61511 is adopted with national variations — as ANSI/ISA 61511 in the United States, and through regional regulatory instruments that impose additional documentation and competence requirements elsewhere — but the technical core of the SIL argument is common to all of them. Table 14 maps each PGDF layer to the clause families it addresses, which is the form in which the framework must be presented to an assessor. The mapping exposes a point that deserves emphasis: Layers 2 to 4 do not participate in the safety instrumented function and therefore do not require safety certification, but they do fall within management of change, competence and operational verification requirements, because they alter the maintenance regime on which the SIL claim depends. A common implementation error is to treat an advisory analytics layer as entirely outside the safety lifecycle; if its output changes when a proof test is performed, it is inside.

Table 14

Mapping of PGDF layers to IEC 61511-1 clause families and the evidence each requires

PGDF layer	IEC 61511-1 clause family	Evidence required	In SIF boundary?
1. Safety-rated hardware	Clause 11 (SIS design and engineering); Clause 12 (application programming)	SIL verification calculation; architectural constraint assessment; systematic capability certificates or prior-use justification	Yes
2. Telemetry acquisition	Clause 11.2 (SIS separation and independence); Clause 8 (process hazard and risk analysis, as modified)	Demonstration of non-interference; one-way interface design review; cyber-security risk assessment	No
3. Prognostic engine	Clause 16 (operation and maintenance); Clause 17 (modification)	Model validation record; performance monitoring plan; retraining and drift-detection procedure	No
4. Maintenance and assurance	Clause 5 (management of functional safety); Clause 16.3 (proof testing); Clause 5.2.6 (competence)	Revised proof-test procedure; competence records; functional safety assessment evidence package	No (procedural)

Note. SIF = safety instrumented function; SIS = safety instrumented system. Clause references follow IEC 61511-1:2016. Layers outside the SIF boundary still fall within management-of-change and competence requirements because they modify the maintenance regime on which the safety integrity level claim depends.

4.8 Relation to prior work

The data-driven process safety literature has advanced rapidly on the detection of process events. Spatial and temporal neural network models have been applied to gas leakage detection with strong reported performance (Kopbayev et al., 2022), engineered features have been used to construct intelligent warning models for natural gas release (Kang et al., 2023), and multimodal fusion of electronic-nose and thermal imaging data has been shown to improve leakage classification (Attallah, 2023; Narkhede et al., 2022), and the robustness of such fusion schemes to degraded input channels has been examined directly (Rahate et al., 2023). Hybrid convolutional and recurrent architectures have been deployed for hazard monitoring in mining environments (Dey et al., 2021; Tripathy et al., 2021). These contributions address a different question from ours: they improve the detection of the hazardous condition, whereas we address the detection of the failure of the instrument charged with detecting it.

The sensor-health literature is closer to our problem. Drift compensation using attention-based recurrent models (Chaudhuri et al., 2021) and surface-state-informed gated recurrent regression (Zhuang et al., 2024) demonstrate that the degradation of gas-sensing elements is learnable, and the selectivity review of Djeziri et al. (2024) surveys the data-driven approaches available. Convolutional autoencoders have achieved real-time sensor fault detection and correction (Jana et al., 2022), gradient-boosted ensembles have been applied to vibration sensor fault diagnosis (Fan et al., 2023), unsupervised methods have been characterized for multivariate industrial time series (Belay et al., 2023), and resource-constrained edge implementations of drift-fault detection have been demonstrated on single-board hardware (Saeed et al., 2020), which is the deployment class assumed for Layer 2 of our framework. The parallel review literature on intelligent fault diagnosis in rotating energy assets documents the same migration from threshold alarms to learned health indicators (Liu et al., 2024). Our contribution relative to this body of work is not a superior algorithm — a random forest on engineered trend features is deliberately unexceptional — but the translation of diagnostic performance into the quantities that determine regulatory compliance, together with an explicit accounting of how much of the resulting risk reduction the analytics actually own.

On the reliability side, dynamic reliability digital twins driven by artificial intelligence have been demonstrated for predictive maintenance of industrial components (D'Urso et al., 2024), and copula-based methods have been developed for systems with dependent failures (Gu et al., 2022), which is the problem class that the β-factor treatment in Equations 2 and 3 approximates. Our treatment of common cause is deliberately conservative relative to these more sophisticated dependency models, and adopting them would be a natural extension.

4.9 Demonstrated mechanisms and hypotheses

Because this is a simulation study, the distinction between what the analysis demonstrates within its own assumptions and what remains conjectural is central to its honest interpretation. Table 15 states that partition explicitly. Claims in the first column follow deductively from the specified models and would be reproduced by any correct re-implementation; claims in the second require empirical field data and should not be relied upon in a safety case until that data exists.

Table 15

Explicit separation of demonstrated mechanisms from hypotheses requiring field validation

Demonstrated within the model assumptions	Hypothesis requiring field validation
Legacy architecture fails both the PFDavg and the architectural-constraint tests for SIL 2, for independent reasons	That the reliability parameters assumed here represent any particular operating facility's installed base
Final-element PFDavg dominates the legacy budget, so detector-only modernization cannot close the gap	That the 72.5% final-element share generalizes to SIFs with different valve populations or existing partial-stroke provision
Prognostic discrimination is high for degradation-mode and absent for shock-mode failures	The degradation-to-shock mixture at a real facility, which sets the achievable benefit ceiling
Effective diagnostic coverage is a second-order influence on PFDavg, ranking seventh of nine parameters behind test-independent failure, common cause and valve failure rate	That real prognostic systems achieve the diagnostic-coverage uplift from 0.88 to 0.96 assumed in Table 3
Covert-failure exposure falls substantially under a condition-based policy given the modelled lead-time distribution	That maintenance organizations sustain the assumed 5–21 day work-order response under real resource constraints
A penalized linear model approaches ensemble discrimination on these features	That linear transparency materially eases functional safety assessment in practice
Precision falls steeply beyond about 60% recall, creating an alarm-burden constraint	That persistence gating, risk ranking or mechanism-aware triage relieve that constraint without unacceptable sensitivity loss
The modelled cost structure yields payback under five years	The valuation of avoided spurious trips, on which the economics materially depend

Note. PFDavg = average probability of failure on demand; SIL = safety integrity level. Entries in the left column are deductive consequences of the specified models and the fixed random seed; entries in the right column are conjectures that the present design cannot test.

4.10 Limitations

The overriding limitation is that no field data were used. Every result is conditional on a data-generating process that we specified, and the degradation model in Equation 5 is almost certainly smoother and more regular than real detector deterioration, which is subject to step changes at maintenance interventions, seasonal effects not captured by a single sinusoid, and operator interactions absent from the simulation. Prognostic performance reported here should

therefore be read as an upper bound on what an equivalent field deployment would achieve for the degradation-mode subpopulation. The competing shock-failure risk was introduced specifically to prevent this optimism from propagating unchecked, but its 30% incidence is itself an assumption rather than an estimate.

Second, statistical inference in a simulation study has a different meaning from inference in an experiment. Sample sizes were chosen by the analyst, so p values can be driven arbitrarily small by extending the simulation, and the extremely small values reported for the deployment comparisons should be read as confirming that the observed differences are not sampling noise rather than as evidence of importance. Effect sizes and interval estimates carry the interpretive weight, and the Hedges' g of 1.34 for covert-failure exposure is the meaningful quantity.

Third, the prior shift between training and test windows — prevalence rising from 2.28% to 20.99% — means that prevalence-dependent metrics are conditional on a mature, degrading population. Average precision, positive predictive value and the Brier score would be materially worse in the early life of a detector fleet. The failure of Platt recalibration to correct this, reported in Section 3.8, indicates that periodic re-estimation rather than one-off calibration is required, and we did not evaluate any re-estimation schedule.

Fourth, the reliability model uses simplified low-demand approximations with a β-factor common-cause treatment and an additive test-independent failure term. It does not model partial-stroke test-induced failures, staggered testing, imperfect repair, or dependency between detector degradation and common-cause susceptibility, all of which would be captured by a Markov or Petri-net formulation. Systematic failures, which some industry analyses hold to dominate real SIS failure experience, are represented only through the qualitative systematic capability gate and the test-independent failure term.

Fifth, cyber-security was addressed only by architectural separation. A telemetry acquisition layer with a read interface to safety-rated devices creates an attack surface that a full deployment must assess formally, and the one-way interface specified in Table 2 is a design intention rather than an evaluated control. Sixth, the economic analysis omits the monetized value of the risk reduction itself, model maintenance and retraining cost, and organizational change cost, and it is sensitive to the valuation of avoided spurious trips. Finally, a single reference facility with one SIF was modelled; generalization to different process units, detector densities, gas species and climatic environments is untested.

4.11 Future work

The most valuable next step is not a better model but a field validation study with a prospectively defined failure-mode partition, in which detector telemetry is collected alongside maintenance records that classify each dangerous failure as degradation- or shock-initiated. That study would replace the single most consequential assumption in this work with an estimate. Beyond it, three extensions follow directly: replacing the analytic PFDavg approximations with a Markov or Petri-net model that represents imperfect repair and staggered testing explicitly; evaluating persistence gating and capacity-constrained risk ranking against the alarm-burden constraint quantified in Section 4.5; and developing a formal systematic capability argument for an advisory machine-learning layer, which the standards do not currently address and which is likely to become the binding obstacle to adoption long before discrimination performance does.

5. Conclusion

A legacy gas detection safety instrumented function documented as SIL 2 was shown, under 120,000-iteration Monte Carlo analysis, to achieve a median PFDavg of 1.73×10^{-2} and to satisfy the SIL 2 requirement in only 4.6% of iterations, while being independently capped at SIL 1 by architectural constraints on its logic solver and final element. The common industrial remedy of detector and logic-solver replacement was insufficient, reducing PFDavg by 19.5% and leaving the mitigated event frequency above the tolerable criterion. The integrated Predictive Gas Detection Framework achieved 1.88×10^{-3}, an 89.0% paired reduction (95% CI 82.4–93.9), satisfied SIL 2 in 100.0% of iterations and met the tolerable frequency with a 3.6-fold margin, at a discounted payback of 4.68 years.

The prognostic layer performed well on the failures it can see and not at all on those it cannot: AUC 0.892 for gradual degradation against 0.545 for abrupt shock failures. It reduced mean covert-failure exposure from 88.2 to 21.1 days (Hedges' g = 1.34, 95% CI 1.08–1.61), averted 75.7% of dangerous undetected failures before onset, and cut spurious alarm burden by 69.0%. It contributed comparatively little to PFDavg itself, where partial-stroke coverage and final-element redundancy dominated.

The practical conclusion is therefore a qualified one, and the qualification is the point. Predictive diagnostics are a legitimate and valuable component of legacy SIS modernization, but they are a complement to hardware integrity rather than a substitute for it, and their contribution is bounded above by the proportion of dangerous failures that proceed through a learnable degradation pathway. Facilities should establish that proportion from their own maintenance records before committing to the architecture, should direct scarce capital to the final element first, and should treat any claim that analytics alone can restore a lapsed SIL with scepticism. All findings derive from simulated data and require field validation before operational adoption.

REFERENCES

1. Attallah, O. (2023). Multitask deep learning-based pipeline for gas leakage detection via e-nose and thermal imaging multimodal fusion. Chemosensors, 11(7), 364.
2. Belay, M. A., Blakseth, S. S., Rasheed, A., & Salvo Rossi, P. (2023). Unsupervised anomaly detection for IoT-based multivariate time series: Existing solutions, performance analysis and future directions. Sensors, 23(5), 2844.
3. Chaudhuri, T., Wu, M., Zhang, Y., Liu, P., & Li, X. (2021). An attention-based deep sequential GRU model for sensor drift compensation. IEEE Sensors Journal, 21(6), 7908–7917.
4. D'Urso, D., Chiacchio, F., Cavalieri, S., Gambadoro, S., & Moheb Khodayee, S. (2024). Predictive maintenance of standalone steel industrial components powered by a dynamic reliability digital twin model with artificial intelligence. Reliability Engineering & System Safety, 243, 109859. https://doi.org/10.1016/j.ress.2023.109859
5. Davidson-Pilon, C. (2019). lifelines: Survival analysis in Python. Journal of Open Source Software, 4(40), 1317.
6. Dey, P., Chaulya, S. K., & Kumar, S. (2021). Hybrid CNN-LSTM and IoT-based coal mine hazards monitoring and prediction system. Process Safety and Environmental Protection, 152, 249–263.
7. Djeziri, M., Benmoussa, S., Bendahan, M., & Seguin, J.-L. (2024). Review on data-driven approaches for improving the selectivity of MOX sensors. Microsystem Technologies, 30(7), 791–807.
8. Fan, C., Li, C., Peng, Y., Shen, Y., Cao, G., & Li, S. (2023). Fault diagnosis of vibration sensors based on triage loss function-improved XGBoost. Electronics, 12(21), 4442.
9. Gu, Y.-K., Fan, C.-J., Liang, L.-Q., & Zhang, J. (2022). Reliability calculation method based on the Copula function for mechanical systems with dependent failure. Annals of Operations Research, 311(1), 99–116.
10. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362.
11. International Electrotechnical Commission. (2010a). IEC 61508-1:2010. Functional safety of electrical/electronic/programmable electronic safety-related systems — Part 1: General requirements (2nd ed.). IEC.
12. International Electrotechnical Commission. (2010b). IEC 61508-2:2010. Functional safety of electrical/electronic/programmable electronic safety-related systems — Part 2: Requirements for electrical/electronic/programmable electronic safety-related systems (2nd ed.). IEC.
13. International Electrotechnical Commission. (2010c). IEC 61508-6:2010. Functional safety of electrical/electronic/programmable electronic safety-related systems — Part 6: Guidelines on the application of IEC 61508-2 and IEC 61508-3 (2nd ed.). IEC.
14. International Electrotechnical Commission. (2016a). IEC 61511-1:2016. Functional safety — Safety instrumented systems for the process industry sector — Part 1: Framework, definitions, system, hardware and application programming requirements (2nd ed.). IEC.
15. International Electrotechnical Commission. (2016b). IEC 61511-2:2016. Functional safety — Safety instrumented systems for the process industry sector — Part 2: Guidelines for the application of IEC 61511-1:2016 (2nd ed.). IEC.
16. International Electrotechnical Commission. (2016c). IEC 61511-3:2016. Functional safety — Safety instrumented systems for the process industry sector — Part 3: Guidance for the determination of the required safety integrity levels (2nd ed.). IEC.
17. International Electrotechnical Commission. (2016d). IEC 60079-29-1:2016. Explosive atmospheres — Part 29-1: Gas detectors — Performance requirements of detectors for flammable gases. IEC.
18. Jana, D., Patil, J., Herkal, S., Nagarajaiah, S., & Duenas-Osorio, L. (2022). CNN and convolutional autoencoder (CAE) based real-time sensor fault detection, localization, and correction. Mechanical Systems and Signal Processing, 169, 108723.

19. Kang, Z., Qian, X., & Li, Y. (2023). Feature extraction of natural gas leakage for an intelligent warning model: A data-driven analysis and modeling. Process Safety and Environmental Protection, 174, 574–584.
20. Kopbayev, A., Khan, F., Yang, M., & Halim, S. Z. (2022). Gas leakage detection using spatial and temporal neural network model. Process Safety and Environmental Protection, 160, 968–975.
21. Liu, X., Chen, Y., Xiong, L., Wang, J., Luo, C., Zhang, L., & Wang, K. (2024). Intelligent fault diagnosis methods toward gas turbine: A review. Chinese Journal of Aeronautics, 37(4), 93–120.
22. Narkhede, P., Walambe, R., Chandel, P., Mandaokar, S., & Kotecha, K. (2022). MultimodalGasData: Multimodal dataset for gas detection and classification. Data, 7(8), 112.
23. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
24. Rahate, A., Mandaokar, S., Chandel, P., Walambe, R., Ramanna, S., & Kotecha, K. (2023). Employing multimodal co-learning to evaluate the robustness of sensor fusion for industry 5.0 tasks. Soft Computing, 27(7), 4139–4155.
25. Saeed, U., Jan, S. U., Lee, Y.-D., & Koo, I. (2020). Machine learning-based real-time sensor drift fault detection using Raspberry Pi. In 2020 International Conference on Electronics, Information, and Communication (pp. 1–7). IEEE.
26. Sarnovsky, M., & Marcinko, J. (2021). Adaptive bagging methods for classification of data streams with concept drift. Acta Polytechnica Hungarica, 18(3), 47–63.
27. Tripathy, D. P., Parida, S., & Khandu, L. (2021). Safety risk assessment and risk prediction in underground coal mines using machine learning techniques. Journal of The Institution of Engineers (India): Series D, 102(2), 495–504.
28. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods, 17(3), 261–272.
29. Zhuang, Y., Yin, D., Wu, L., Niu, G., & Wang, F. (2024). A deep learning approach for gas sensor data regression: Incorporating surface state model and GRU-based model. APL Machine Learning, 2(1), 016104. https://doi.org/10.1063/5.0160983