

American Journal of Scholarly Research and Innovation

Volume: *14;* ***Issue:*** *12*
Pages: *01– 40*
eISSN: 3067-2163

Uncertainty-Calibrated Explainable AI for Cybersecurity Alert Prioritization: Improving Detection Reliability and Human Decision-Making

Maishar Mostari Zenia[1]

[1]. Graduate Student, M.S. in Cybersecurity, Southeast Missouri State University, Cape Girardeau, Missouri, USA; Email: mzenia1s@semo.edu

Doi: 10.63125/an0pcq39
Received: *21 September 2024;* ***Revised:*** *24 October 2024; Accepted: 12 November 2024; Published: 02 December 2024;*

Abstract

Security operations centres receive far more alerts than analysts can review, and the probabilities, rankings and explanations that machine learning attaches to those alerts are rarely checked for reliability. This article designs a data-driven analytics framework that makes alert scores mean what they say, carries uncertainty alongside them, and explains them in a way whose quality is measured, and it evaluates the framework on a transparent synthetic testbed: 1,200 alerts per day from eight families, a rare and drifting incident class, four asset-value tiers, a fixed review capacity of 12.5%, and a stylised analyst model. Raw scores from deep boosted trees ranked incidents well but under-predicted risk by a factor of five (mean predicted probability 0.84% against a true rate of 4.2%); every post-hoc calibrator cut the expected calibration error from 0.034 to between 0.006 and 0.009, drift raised it to 0.017, and a sliding-window recalibrator using delayed labels held it at 0.008. Calibration did not change a ranking by probability alone, but it changed the queue once probability was multiplied by asset loss: loss captured rose from 59.7% to 68.3%, Tier-3 incidents missed fell from 27.6% to 6.9%, and net value rose from 29.4 to 34.5 thousand USD per day. Conformal sets, ensemble spread and explanations behaved less favourably: spread did not signal drift, explanation stability (0.65) was unrelated to uncertainty, and in the analyst model uncertainty flags and explanations traded recall against precision without raising net value. In a 2,000-draw Monte Carlo analysis the calibrated queue beat the raw-score baseline in every draw, and Sobol indices attribute the gain mainly to loss scale and attack prevalence. All results describe the testbed, and the analyst findings are hypotheses for studies with analysts, not findings about them. A governance record and a four-phase roadmap are proposed.

Keywords

Alert Prioritization, Security Operations Centre, Probability Calibration, Uncertainty Quantification, Conformal Prediction, Explainable AI, Human-AI Decision-Making, Alert Fatigue, Cost-Sensitive Triage, Decision Analytics

Introduction

A security operations centre (SOC) is a queueing system with an adversary. Detectors, rules and analytics emit alerts; analysts investigate as many as their time allows; and the incidents that matter must be found among a stream in which most alerts are false. The imbalance is old. Axelsson showed that when attacks are rare, even a detector with a low false-positive rate produces mostly false alarms (Axelsson, 2000), and studies of working analysts describe alert volumes, false positives and tool sprawl as central sources of frustration and burnout (Alahmadi et al., 2022; Kokulu et al., 2019; Sundaramurthy et al., 2015; Vielberth et al., 2020). Machine learning is the usual response. Classifiers, clustering and provenance analysis reduce false positives and rank what remains (Julisch, 2003; Pietraszek, 2004; Hassan et al., 2019; Veeramachaneni et al., 2016), and explainable AI (XAI) is added so that the analyst can see why an alert scored highly (Ribeiro et al., 2016; Lundberg & Lee, 2017; Capuano et al., 2022; Nadeem et al., 2023).

What the analyst sees, however, is usually a number, a rank and an explanation, and each is a claim that can be checked. The number is read as a probability, but classifier scores are often not probabilities: modern classifiers are miscalibrated, boosted trees distort the scores they emit, and calibration degrades when the data shift (Guo et al., 2017; Niculescu-Mizil & Caruana, 2005; Ovadia et al., 2019; Minderer et al., 2021). A rank hides how sure the model is, and an alert the model has never seen anything like looks the same as a routine one. An explanation is a second model of the first, and attributions can be unstable, unfaithful or manipulable (Adebayo et al., 2018; Ghorbani et al., 2019; Slack et al., 2020; Krishna et al., 2022). Meanwhile the human and the machine form a team whose performance depends on how much the analyst relies on the machine; reliance that is too high produces automation bias and complacency, and reliance that is too low wastes the machine (Parasuraman & Riley, 1997; Parasuraman & Manzey, 2010; Lee & See, 2004; Skitka et al., 1999). Calibrated probabilities, honest uncertainty and measured explanations are the machine's side of that bargain.

This article asks what an analytics framework for alert prioritization should contain if the probability, the uncertainty and the explanation are all to be reliable, and what such a framework is worth. Four research questions guide it. RQ1: how miscalibrated are alert scores, how much do calibrators repair, and how do the repairs hold up under drift, across alert families and with few labels? RQ2: what do uncertainty signals (ensemble spread and conformal sets) add, in coverage, abstention and error? RQ3: in a capacity-limited queue with asset values, what is the decision value of calibrated and uncertainty-aware prioritization? RQ4: how do explanations behave, and what, under an explicit and stylised analyst model, do calibrated displays, uncertainty flags and explanations do to decisions and to net value, and which assumptions matter most?

The method is design science with a simulation study (Hevner et al., 2004; Peffers et al., 2007; Gregor & Hevner, 2013): requirements are drawn from the literature, a four-layer framework is designed from them, and the framework is exercised on a transparent synthetic testbed whose every parameter is stated, because no real SOC data were used (Ashik & Shabbir, 2022; Hasan & Adnan, 2022; Imran & Jannatul, 2021). The testbed has eight alert families, a rare incident class with a hard subset, four asset-criticality tiers, a covariate and prior shift in the test period, delayed labels, and a review capacity of 12.5% of daily alerts (Akter & Sumaiya, 2022; Mashfiqur & Md, 2022; Md & Yeasin, 2022). Five detectors, six calibrators, class-conditional conformal sets, ensemble spread, Shapley attributions and a noisy-Bayesian analyst model are evaluated with three seeds, and the decision value is propagated through 2,000 Monte Carlo draws with Sobol indices and a metamodel (Sobol', 2001; Saltelli et al., 2010).

The contribution is threefold. It states six requirements for reliable, explainable alert prioritization and a framework that meets them. It reports, on one transparent testbed, where calibration changes decisions (when probability is combined with value or compared with a cost) and where it does not (when a single score is merely ranked), and where uncertainty signals and explanations disappoint. And it provides a governance record, a set of reliance and calibration metrics, and a phased roadmap for a SOC that wants to test the ideas on its own alerts. The findings are about the testbed. The orderings, rather than the numbers, are what a reader should carry to other settings, and the analyst results in particular are hypotheses to be tested with people.

Background and Related Work

Alert fatigue and the work of the SOC

The practice literature on SOCs is mostly qualitative. Interviews with analysts and managers describe a work system in which tools generate more alerts than staff can handle, in which false positives dominate, and in which the metrics used to manage the centre (volume closed, time to close) do not reward careful investigation (Kokulu et al., 2019; Alahmadi et al., 2022). A human-capital model of analyst burnout treats alert load and monotony as drivers of turnover (Sundaramurthy et al., 2015), and a systematic study of SOC research lists people, process and technology challenges, among them the integration of analytics with analyst workflow (Vielberth et al., 2020). Parallel evidence from health care, where alarm fatigue and alert override are well documented, shows how repeated, mostly irrelevant alerts erode attention and trust (Cvach, 2012; Ancker et al., 2017), and the old warning of Breznitz about the psychology of false alarms describes the same mechanism (Breznitz, 1984). The base-rate arithmetic is the quantitative core: when incidents are a few per cent of alerts, precision is low unless discrimination is very high (Axelsson, 2000; Sommer & Paxson, 2010; Arp et al., 2022). Burnout is also a cost to the organisation, and the cost of the queue is therefore not only the cost of the time spent but the cost of the attention lost (Maslach et al., 2001). These sources describe perceptions and organisational patterns; they do not provide measurements of the reliability of the numbers on the analyst's screen, which is the gap this article addresses.

Alert correlation, triage and learning

Reducing and ranking alerts has a long history. Alert correlation merges related alarms into attack scenarios (Cuppens & Miège, 2002; Valdes & Skinner, 2001); clustering finds root causes that generate most false alarms (Julisch & Dacier, 2002; Julisch, 2003); adaptive classifiers learn from analyst feedback to suppress false positives (Pietraszek, 2004; Spathoulas & Katsikas, 2010); belief propagation over enterprise logs ranks suspicious early-stage infections (Oprea et al., 2015); analyst-in-the-loop systems combine unsupervised and supervised models and learn from analyst labels (Veeramachaneni et al., 2016); and provenance-based triage scores alerts by the rarity of the causal paths behind them (Hassan et al., 2019; Hassan et al., 2020). Evaluations of intrusion detection have themselves been criticised for unrepresentative data and experimental bias (McHugh, 2000; Tavallaee et al., 2009; Engelen et al., 2021; Arp et al., 2022). Most of this work reports ranking quality, precision and recall. It seldom asks whether the scores are probabilities, which is what an expected-value queue, a cost-based threshold or a trusting analyst needs.

Probability calibration

A classifier is calibrated when, among the cases it scores at 0.2, about one in five is positive. The concept is old in forecasting, where reliability diagrams and decompositions of the Brier score were developed (Brier, 1950; Murphy, 1973; DeGroot & Fienberg, 1983; Dawid, 1982; Bröcker & Smith, 2007), and proper scoring rules explain why the Brier score and log loss reward calibrated, sharp forecasts (Gneiting & Raftery, 2007). For classifiers, Platt scaling and isotonic regression map scores to probabilities (Platt, 2000; Zadrozny & Elkan, 2002), and Niculescu-Mizil and Caruana showed that boosted trees need such mapping because boosting pushes predicted probabilities away from the extremes (Niculescu-Mizil & Caruana, 2005). Beta calibration and Dirichlet calibration extend the family (Kull et al., 2017; Kull et al., 2019), and temperature scaling, a single-parameter rescaling of the logit, repairs much of the miscalibration of neural networks (Guo et al., 2017). Expected calibration error (ECE) is the usual summary, but binned estimators are biased and sensitive to binning, and calibration needs tests and not only a number (Naeini et al., 2015; Nixon et al., 2019; Vaicenavicius et al., 2019; Widmann et al., 2019; Kumar et al., 2019). Newer architectures are not uniformly worse calibrated than older ones (Minderer et al., 2021), and per-group calibration (multicalibration) can fail even when overall calibration holds (Hébert-Johnson et al., 2018; Pleiss et al., 2017). For security, the relevant questions are practical: how much labelled data a calibrator needs, whether calibration survives drift, and whether one calibrator serves all alert families.

Uncertainty estimation

Predictive uncertainty has an aleatoric part (noise in the data that no model can remove) and an epistemic part (ignorance that more data or a better model could reduce) (Kendall & Gal, 2017; Hüllermeier & Waegeman, 2021; Depeweg et al., 2018). Deep ensembles, Monte Carlo dropout,

Bayesian approximations and evidential methods estimate it (Lakshminarayanan et al., 2017; Gal & Ghahramani, 2016; Maddox et al., 2019; Sensoy et al., 2018; Wilson & Izmailov, 2020; Fort et al., 2019), and surveys compare them (Abdar et al., 2021; Gawlikowski et al., 2023). Their behaviour under dataset shift is the test that matters for security, and the evidence is that quality degrades with shift and that ensembles are a strong baseline (Ovadia et al., 2019). The maximum softmax probability is a simple baseline for flagging misclassified and out-of-distribution inputs (Hendrycks & Gimpel, 2017). Whether the spread of an ensemble rises on drifted security alerts, and whether it distinguishes hard incidents from easy ones, is an empirical question that the testbed can answer for itself (Ashfaq & Mohammad, 2022; Imran & Jannatul, 2022; Morshedul & Islam, 2022).

Conformal prediction and risk control

Conformal prediction wraps any predictor in sets with a finite-sample coverage guarantee under exchangeability (Vovk et al., 2005; Shafer & Vovk, 2008; Angelopoulos & Bates, 2023). Split (inductive) conformal prediction uses a held-out calibration set (Papadopoulos et al., 2002; Lei et al., 2018), class-conditional or Mondrian variants give coverage within each class, which matters when one class is rare (Vovk et al., 2005; Vovk, 2012), and adaptive or weighted variants address covariate shift and non-exchangeable data (Tibshirani et al., 2019; Gibbs & Candès, 2021; Barber et al., 2023). Set-valued classifiers, adaptive sets and risk-controlling sets extend the idea (Sadinle et al., 2019; Romano et al., 2020; Angelopoulos et al., 2021; Bates et al., 2021; Angelopoulos et al., 2024). Conditional coverage cannot be guaranteed in distribution-free settings (Foygel Barber et al., 2021). In malware classification, conformal evaluation has been used to detect concept drift (Jordaney et al., 2017; Barbero et al., 2022). For alert triage, a conformal set with both classes in it is a principled version of "the model does not know", and this article uses it as an escalation signal.

Selective prediction, deferral and human-AI teams

A classifier with a reject option abstains when its confidence is low, and the trade-off between coverage and risk goes back to Chow (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; Geifman & El-Yaniv, 2019). Learning to defer extends abstention by training the model jointly with the decision about which cases a human should handle (Madras et al., 2018; Mozannar & Sontag, 2020; Okati et al., 2021; Raghu et al., 2019), and complementarity is the goal: the team should do better than either member (Wilder et al., 2020; Bansal et al., 2019; Kamar et al., 2012; Horvitz, 1999). Empirically, human decisions can be improved by machine predictions in some settings (Kleinberg et al., 2018), but the labels on which such systems are trained and evaluated are selective, because outcomes are observed only for cases that humans acted on (Lakkaraju et al., 2017). That problem is acute in a SOC, where unreviewed alerts are rarely labelled.

Dataset shift and drift in security

Security data drift because adversaries adapt, software changes and the mix of traffic changes (Kaniz & Jahanara, 2022; Sadia & Mostari, 2022; Stabak & Shahid, 2022). Sommer and Paxson explained why machine learning for intrusion detection is hard, among other reasons because of the open-world problem and the cost of errors (Sommer & Paxson, 2010); Arp and colleagues catalogued pitfalls in learning-based security research (Arp et al., 2022); TESSERACT showed how temporal and spatial bias inflate malware-classifier results (Pendlebury et al., 2019); and concept drift in classifiers is a standing problem (Gama et al., 2014; Lu et al., 2019; Quiñonero-Candela et al., 2009; Moreno-Torres et al., 2012). Label-shift methods adjust probabilities when class priors change (Saerens et al., 2002; Lipton, 2018). For calibrated alert probabilities, drift is a first-order concern: a calibrator fitted on last quarter's alerts may be wrong this week, and the corrections available (recalibration on recent labelled alerts, prior correction, adaptive conformal updates) depend on labels that arrive late (Mohammad, 2023; Adnan & Iftekhar, 2023; Feroz & Pervaz, 2022).

Explainable AI for security and its quality

Post-hoc explanation methods attribute a prediction to features, through local surrogates, anchors, Shapley values or counterfactuals (Ribeiro et al., 2016; Ribeiro et al., 2018; Lundberg & Lee, 2017; Štrumbelj & Kononenko, 2014; Mothilal et al., 2020), and there are surveys, taxonomies and critiques (Guidotti et al., 2018; Barredo Arrieta et al., 2020; Adadi & Berrada, 2018; Miller, 2019; Rudin, 2019; Doshi-Velez & Kim, 2017; Lipton, 2018). Shapley values have a clean axiomatic basis (Shapley, 1953), efficient algorithms for trees (Lundberg et al., 2020), and well-known limitations when features are

dependent or when used as importance measures (Kumar et al., 2020; Janzing et al., 2020; Aas et al., 2021; Covert et al., 2021; Fisher et al., 2019). Surveys of explainable AI in cybersecurity and of explainable intrusion detection identify the analyst as the intended reader and the lack of evaluation with analysts as a gap (Capuano et al., 2022; Nadeem et al., 2023; Neupane et al., 2022; Mahbooba et al., 2021; Wang et al., 2020). The quality of an explanation is not given. Saliency maps can fail sanity checks, attributions can be changed by small perturbations, explanation methods disagree, and classifiers can be built to fool LIME and SHAP (Adebayo et al., 2018; Ghorbani et al., 2019; Alvarez-Melis & Jaakkola, 2018; Krishna et al., 2022; Slack et al., 2020; Hooker et al., 2019; Yeh et al., 2019; Bhatt et al., 2020). Uncertainty has been proposed as a form of transparency, and explanations can themselves carry uncertainty (Bhatt et al., 2021; Slack et al., 2021; Zhao et al., 2021), which makes attribution spread and stability natural diagnostics.

Human reliance on automation and the effect of explanations

The human-factors literature treats reliance as a decision that should track the automation's reliability (Ahmed & Sadia, 2023; Islam & Md, 2023; Asheke, 2023): the vocabulary of use, misuse, disuse and abuse; levels of automation; complacency and automation bias; trust calibration; and the ironies of automating the easy part and leaving the human the hard part (Parasuraman & Riley, 1997; Parasuraman et al., 2000; Parasuraman & Manzey, 2010; Lee & See, 2004; Hoff & Bashir, 2015; Skitka et al., 1999; Bainbridge, 1983; Cummings, 2004; Goddard et al., 2012). Algorithm aversion and algorithm appreciation show that reliance can err in both directions (Dietvorst et al., 2015; Logg et al., 2019), and situation awareness frames what the operator needs to keep (Endsley, 1995). Studies with people suggest that explanations do not reliably improve joint performance and can increase reliance on incorrect advice (Gazi Yeasin, 2023; Jahanara & Kaniz, 2023; Sheak & Tahmedur, 2023), that displaying a model's confidence can help align trust with correctness, that forcing deliberation can reduce overreliance, and that stated accuracy shapes trust (Bansal et al., 2021; Zhang et al., 2020; Buçinca et al., 2021; Lai & Tan, 2019; Poursabzi-Sangdeh et al., 2021; Vasconcelos et al., 2023; Wang & Yin, 2021; Yin et al., 2019; Kocielnik et al., 2019; Green & Chen, 2019). These are findings about particular tasks and populations, mostly not security analysts, and they are used here only to motivate the analyst model's structure and to flag what must be tested.

Cost-sensitive decisions, queues and governance

Decisions under asymmetric costs have a classical form: act when the expected loss avoided exceeds the cost of acting, which for a binary decision is a threshold on the probability that depends on the costs (Elkan, 2001; Provost & Fawcett, 2001). Threshold choice, cost curves, decision curves and expected-loss views of performance metrics make the dependence on costs and prevalence explicit (Adams & Hand, 1999; Drummond & Holte, 2006; Vickers & Elkin, 2006; Hand, 2009; Hernández-Orallo et al., 2012). When capacity binds, the problem is one of scheduling: ordering jobs by weight over processing time minimises weighted completion on a single resource (Smith, 1956; Pinedo, 2016), and priority queues formalise waiting (Cobham, 1954; Harchol-Balter, 2013). In alert triage, the weight is the expected loss avoided, which is the calibrated probability times the asset's value, and the processing time is the review time (Kaium, 2023; Pervaz, 2023; Badrul & Mominul, 2023). Governance adds records and oversight: the NIST AI Risk Management Framework, ISO/IEC 42001 and 23894 for AI management and risk, ISO/IEC 27035 and NIST SP 800-61 for incident management, and the EU AI Act, whose human-oversight provisions apply to high-risk systems (Tabassi, 2023; ISO/IEC, 2023a; ISO/IEC, 2023c; ISO/IEC, 2023b; Cichonski et al., 2012; European Parliament and Council, 2024b; National Institute of Standards and Technology, 2024). NIST SP 800-61 Revision 2 is cited here; I believe a later revision has since been issued, so the current edition should be checked before the guidance is applied (Nishat, 2023; Sadia, 2023; Stabak & Md Siam, 2023). Whether an alert-prioritization tool falls under the EU AI Act's high-risk categories depends on its use, and legal advice, not this article, should settle it; application dates under the Act are phased and may have changed (Mohammad, 2024; Feroz, 2023; Tahmedur & Rifat, 2023).

Requirements and the framework

Table 1 organises the literature into eleven strands and the gap each leaves. Six requirements follow (Table 2). R1, probabilities that mean what they say: the score shown to the analyst and used in the queue must be calibrated on recent labelled data, and calibration must be reported by alert family and

period. R2, uncertainty beyond the point estimate: the framework must carry a measure of what the model does not know and a rule for escalating alerts in that region. R3, adaptation under drift with delayed labels: calibration and coverage must be monitored and corrected as the data change, using labels that arrive late and only for reviewed alerts. R4, decisions priced in money and capacity: the queue must be ordered by expected loss avoided under a fixed capacity, and thresholds must follow costs. R5, explanations whose quality is measured: attributions must be accompanied by stability, faithfulness and agreement diagnostics, and by a readable surrogate. R6, human use, evidence and governance: reliance, accuracy and time must be measured, and the model version, calibration date, coverage, drift alarms and oversight decisions must be recorded.

Table 1: Literature strands, what they establish, representative sources and the gap addressed

Strand	What the literature establishes	Representative sources	Gap addressed here
SOC operations and alert fatigue	Analysts face very high volumes of alerts, most of them false; fatigue and burnout degrade review; the base-rate fallacy limits detector precision	Alahmadi et al. (2022); Kokulu et al. (2019); Sundaramurthy et al. (2015); Vielberth et al. (2020); Axelsson (2000)	Few studies treat the probability shown to the analyst as a calibrated quantity with a measurable decision value
Alert correlation and triage by learning	Clustering, adaptive classification and provenance-based triage reduce false positives; analyst-in-the-loop learning	Julisch (2003); Pietraszek (2004); Spathoulas & Katsikas (2010); Hassan et al. (2019); Veeramachaneni et al. (2016); Oprea et al. (2015)	Ranking quality is reported; probability quality and its effect on cost-based ranking are not
Probability calibration	Modern classifiers are often miscalibrated; post-hoc calibrators (Platt, isotonic, temperature, beta, Dirichlet) and metrics (ECE, Brier, reliability diagrams)	Platt (2000); Zadrozny & Elkan (2002); Niculescu-Mizil & Caruana (2005); Guo et al. (2017); Kull et al. (2017, 2019); Nixon et al. (2019); Minderer et al. (2021)	Evidence on rare-event alert streams with drift, per-family calibration and calibration-set size is thin
Uncertainty estimation	Ensembles, dropout, evidential and Bayesian methods; aleatoric versus epistemic uncertainty; behaviour under dataset shift	Gal & Ghahramani (2016); Kendall & Gal (2017); Lakshminarayanan et al. (2017); Ovadia et al. (2019); Hüllermeier & Waegeman (2021); Gawlikowski et al. (2023)	Whether ensemble spread flags drift or hard incidents in security alerts is rarely tested
Explainable AI for security	Shapley, LIME, anchors, counterfactuals, surrogates; surveys of explainable intrusion detection and security ML	Ribeiro et al. (2016); Lundberg & Lee (2017); Štrumbelj & Kononenko (2014); Capuano et al. (2022); Nadeem et al. (2023); Neupane et al. (2022)	Explanation quality is rarely related to predictive uncertainty or to analyst outcomes
Explanation quality, fragility and uncertainty	Saliency sanity checks; fragility and manipulation; disagreement among methods; uncertainty as transparency	Adebayo et al. (2018); Ghorbani et al. (2019); Slack et al. (2020, 2021); Bhatt et al. (2021); Krishna et al. (2022); Kumar et al. (2020)	Stability and faithfulness are not reported by confidence level or under drift
Human-automation interaction	Trust, reliance, automation bias, complacency; explanations can raise overreliance; confidence displays can calibrate trust	Parasuraman & Riley (1997); Lee & See (2004); Parasuraman & Manzey (2010); Zhang et al. (2020); Bansal et al. (2021); Buçinca et al. (2021); Green & Chen (2019)	No integrated model linking calibration, flags and explanations to queue-level value for SOC analysts
Decision analytics and governance	Cost-sensitive decisions, decision curves, scheduling rules; AI risk management and incident-handling standards	Elkan (2001); Vickers & Elkin (2006); Hernández-Orallo et al. (2012); Smith (1956); NIST AI RMF (Tabassi, 2023); ISO/IEC (2023a); Cichonski et al. (2012)	Translation of calibration into expected-value queues and governance records

Table 2: Requirements for reliable, explainable alert prioritization: rationale, operationalisation in the framework and where each is evidenced in the testbed

Requirement	Why it matters	How it is operationalised here	Evidence in the testbed
R1 Probabilities that mean what they say	Expected-value ranking, thresholds and analyst trust all use the number as a probability	Post-hoc calibrators fitted on a recent labelled window; reliability curves, ECE, Brier, log loss by family and period	Table 7 and Figures 5 and 6: ECE falls from 0.034 to about 0.006 to 0.009
R2 Uncertainty beyond the point estimate	Hard and shifted alerts need a different treatment from easy ones	Ensemble spread, $p(1 - p)$, class-conditional conformal sets with an ambiguous outcome	Table 9 and Figure 7
R3 Adaptation under drift with delayed labels	Calibration decays when attack mix and detectors change; labels arrive late	Sliding-window recalibration and adaptive conformal update using labels delayed by two days	Table 7 (adaptive Platt) and Figure 7(a)
R4 Decisions priced in money and capacity	A ranking that ignores asset value or capacity misallocates scarce analyst time	Queue ordered by calibrated p times asset loss at a fixed daily capacity; cost-optimal thresholds	Tables 10 and 11 and Figure 8
R5 Explanations whose quality is measured	Attributions can be unstable, unfaithful or manipulated, and may raise overreliance	Shapley values per ensemble member; stability, deletion faithfulness, ground-truth agreement, surrogate fidelity	Table 12 and Figure 9
R6 Human use, evidence and governance	Value is realised only if analysts weigh the machine appropriately; regulators ask for oversight records	Analyst model with nine conditions; reliance metrics; governance record (version, calibration date, coverage, drift alarms)	Tables 13 and 14 and Figures 10 to 12

From these the framework (Figure 1) is designed in five layers: the alert and telemetry layer, the detection layer, the calibration and uncertainty layer, the explanation and prioritisation layer, and the human decision and governance layer. Feedback flows from analyst decisions to labels, from labels to recalibration, and from the governance record to retraining. Figure 2 follows a single alert from detector to analyst and marks where uncertainty enters and how it should shape reliance. The Methods give the testbed and the models.

Figure 1: Uncertainty-calibrated explainable alert-prioritization framework: alert and telemetry, detection, calibration and uncertainty, explanation and prioritisation, and human decision and governance layers, with feedback loops for delayed labels, recalibration and retraining

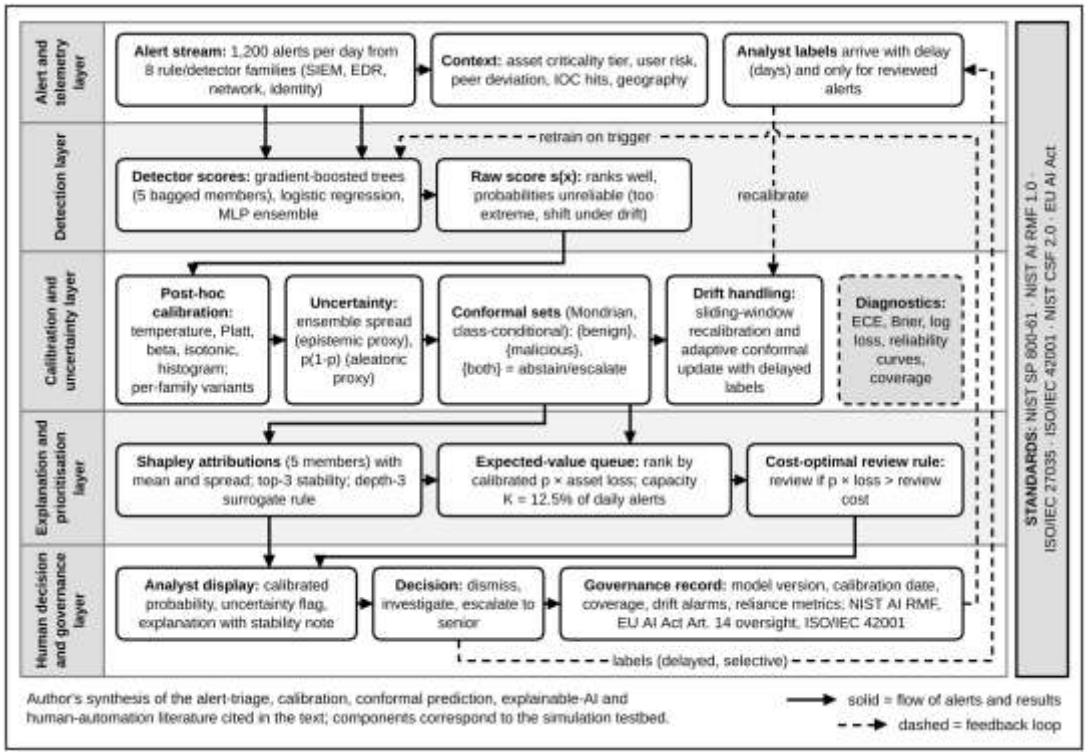
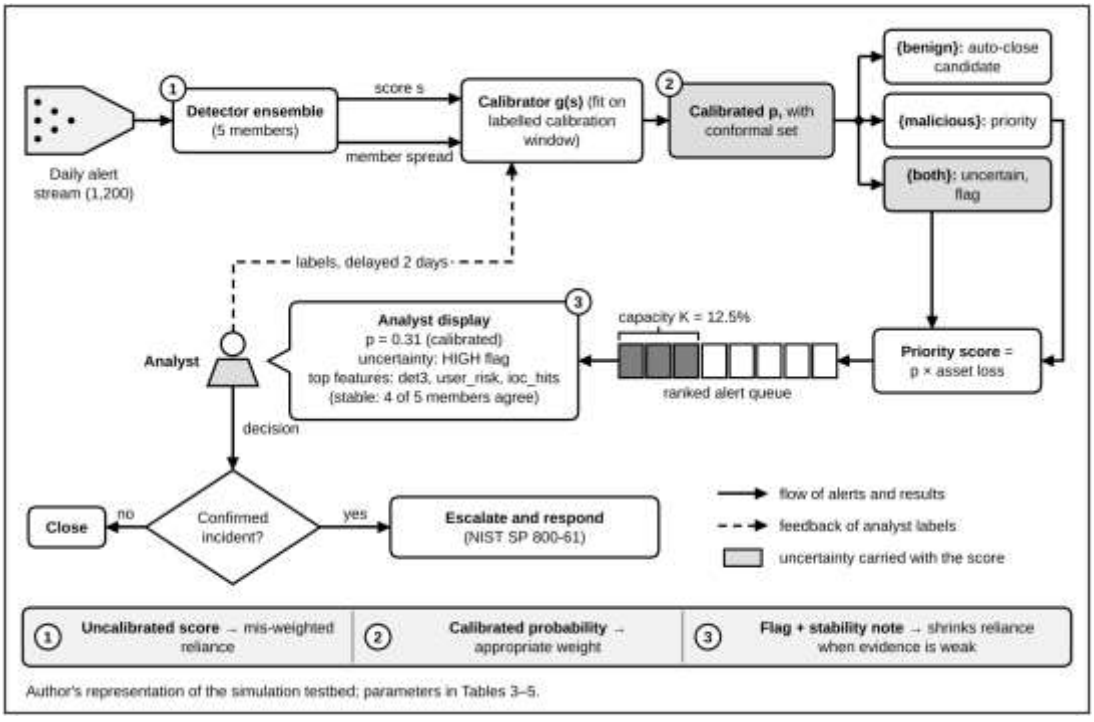


Figure 2: From alert to analyst decision: detector ensemble, calibrator, conformal set, expected-value queue under capacity, analyst display with probability, uncertainty flag and explanation stability, and the delayed-label feedback loop

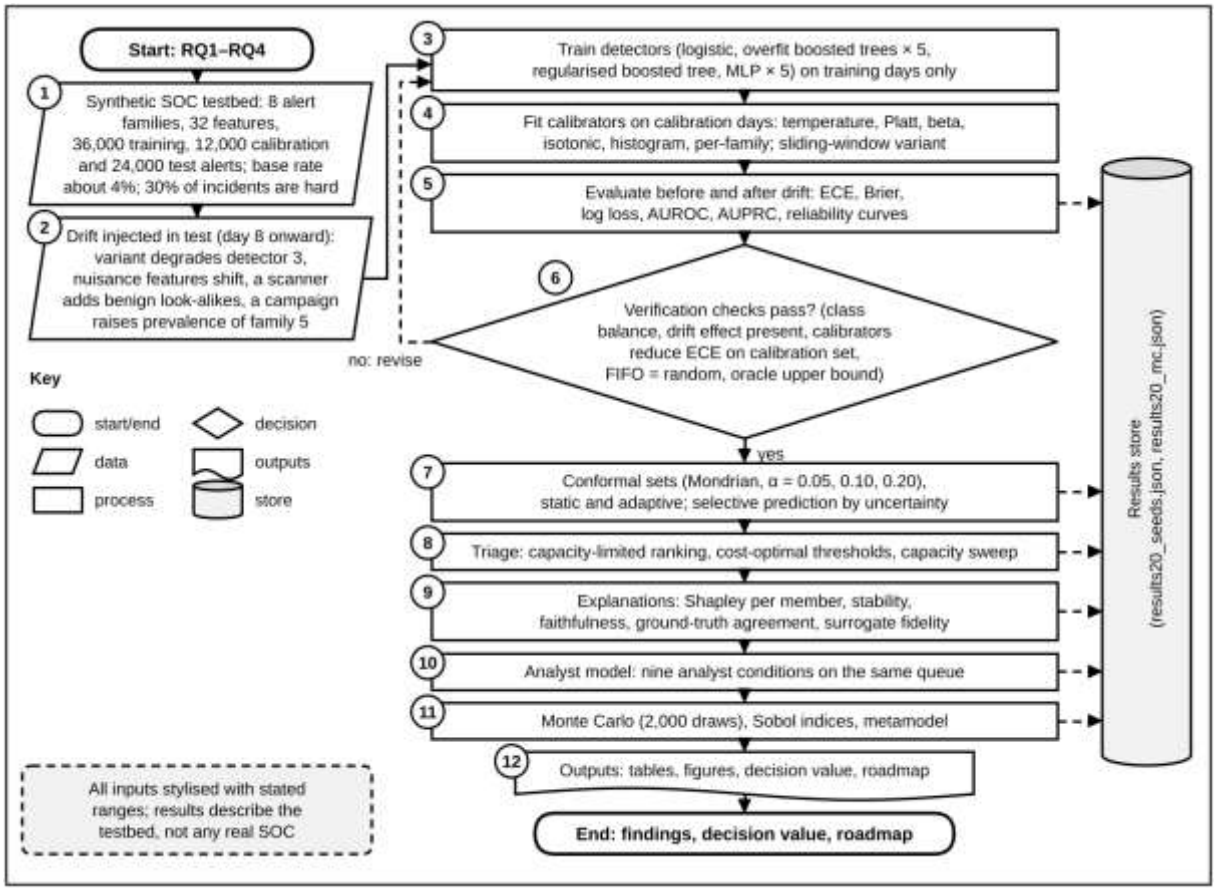


Methods

Study design

The study follows the design-science logic of Hevner and colleagues: an artefact (the framework) is built from requirements and evaluated against stated criteria (Hevner et al., 2004; Peffers et al., 2007; Gregor & Hevner, 2013). Evaluation is by simulation, because the quantities of interest (calibration error under a known drift, the value of a queue ordering, the effect of a display on a decision) need ground truth that a real SOC does not have. Simulation conventions follow Law and Sargent: common random numbers, verification checks, a results file from which every table and figure is produced, and honest statements of validity (Law, 2015; Sargent, 2013). Figure 3 gives the design: a synthetic alert testbed with drift; five detectors and six calibrators; verification; experiments on calibration, conformal sets, triage, explanations and analyst conditions; and a Monte Carlo and Sobol analysis of the decision value. The four research questions map to the experiments: RQ1 to the calibration experiments, RQ2 to conformal sets and spread, RQ3 to the triage experiments and RQ4 to the explanation, analyst and Monte Carlo experiments. Every random element is seeded, three seeds are run, and results are reported as means with standard deviations over seeds.

Figure 3: Study design as a flowchart: synthetic testbed with drift, detectors and calibrators, verification, conformal and triage experiments, explanation and analyst experiments, Monte Carlo and Sobol analysis, outputs



Testbed: alerts, drift, labels and value

The testbed generates 1,200 alerts per day from eight families (Table 3). Days 0 to 29 train the detectors, days 30 to 39 are the calibration window, and days 40 to 59 are the test period, which contains a drift episode from its ninth day. Each alert has 32 features: a family indicator, eight detector scores, eight context features (user risk, peer deviation, asset criticality, hour of day, indicators-of-compromise hits, geographic rarity, volume z-score and one spare) and eight nuisance features. Noise is Student t with five degrees of freedom, scaled by family. Incidents occur at family-specific rates between 1.2% and 10%, for a mean of about 4%; they raise their own family's detector score, user risk, peer deviation and

indicator hits; and 30% of incidents are hard (low-and-slow), with signals about a quarter as strong, so that part of the uncertainty is irreducible. Six per cent of benign alerts are look-alikes that raise their family's detector. Labels are complete and correct in every experiment; the selective, delayed and noisy labels of real SOCs are discussed as a limitation and not simulated (Figure 4(b), (c) and (d) show the prevalence by day, one detector's shift and the asset tiers).

The drift is the kind security teams meet: a variant lowers detector 3's response to family 3 incidents, a new scanner makes benign alerts look like family 0 incidents, nuisance features shift, and a seven-day campaign multiplies the prevalence of family 5. Asset criticality is a feature, observable to the model and the policy, and sets the loss avoided if a true incident is investigated the same day (Table 4). The value of an alert to the queue is therefore the probability of an incident times the asset's loss, and the cost of reviewing it is analyst time. Both the loss scale and the review cost are stylised and are varied in the Monte Carlo analysis. Labels for day d become available on day d + 2, which is the delay used by the adaptive methods.

Figure 4: Testbed: (a) alert families by share of alerts and incident rate, (b) incident prevalence by test day with the drift onset, (c) the shift of one detector's score for family 3 incidents, (d) asset criticality tiers and the loss if a true incident is missed

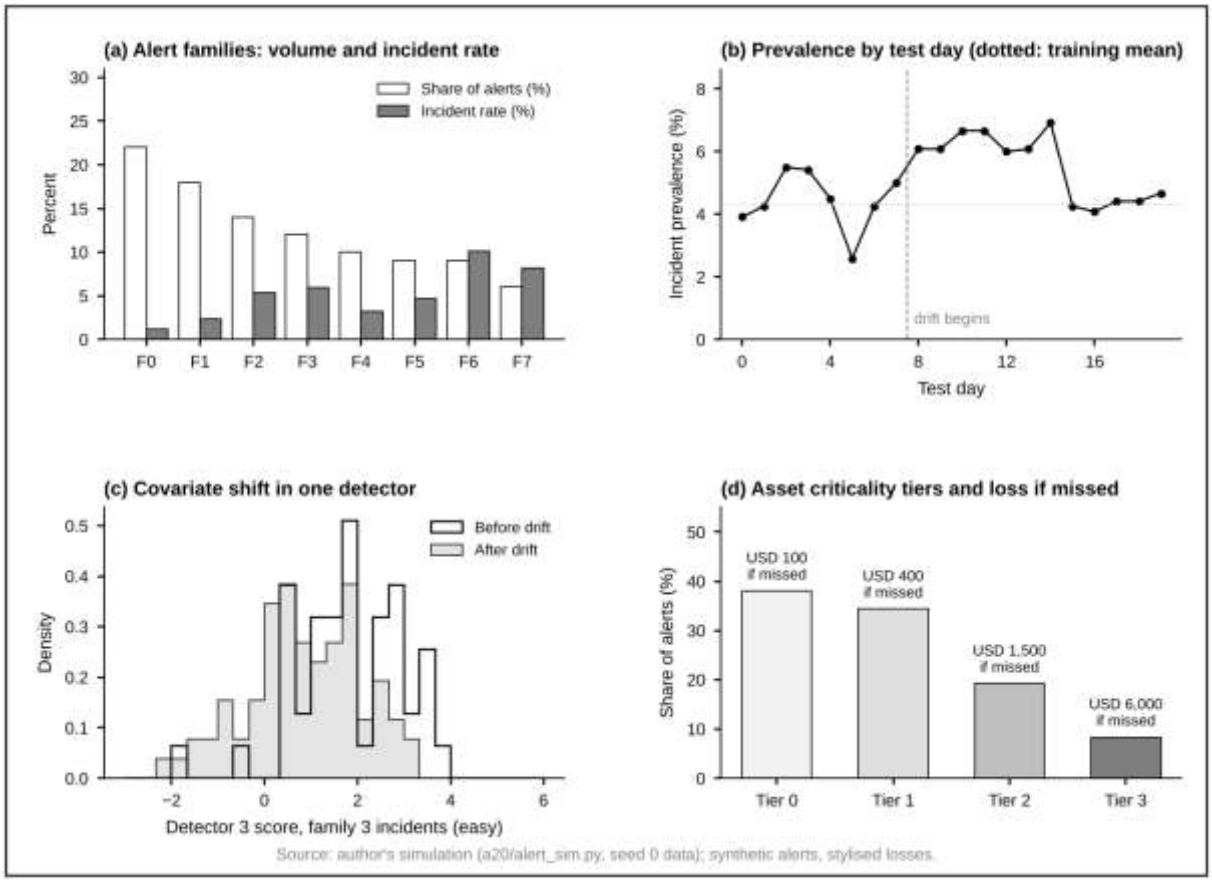


Table 3: Testbed: alert stream, family mix, features, incident signal, hard incidents, drift and seeds

Element	Model	Values (stylised)	Basis
Alert stream	Eight families (rule and detector sources); 1,200 alerts per day; training days 0 to 29, calibration days 30 to 39, test days 40 to 59	36,000 training, 12,000 calibration, 24,000 test alerts; mean incident rate 4.2%	Sommer & Paxson (2010); Alahmadi et al. (2022)
Family mix and incident rates	Shares of alerts and per-family incident rates differ; some families are rare but dangerous	Shares 22, 18, 14, 12, 10, 9, 9, 6% ; rates 1.2, 2.0, 5.0, 6.0, 3.0, 4.5, 10.0, 8.0%	Axelsson (2000)
Features (32)	Family one-hot (8); detector scores (8); context (8): user risk, peer deviation, asset criticality, hour, IOC hits, geographic rarity, volume z-score, one spare; nuisance (8)	Student t noise (5 d.f.); family-specific scale	Chandola et al. (2009)
Incident signal	Incidents raise the family's own detector, user risk, peer deviation and IOC hits; benign look-alikes (6%) also raise the detector	Easy incidents shift the own detector by 1.9; hard incidents by 0.45	Arp et al. (2022)
Hard incidents	A share of incidents are low-and-slow with weak signals everywhere	30% of incidents	Sommer & Paxson (2010)
Drift (test day 8 onward)	Variant degrades detector 3 for family 3 incidents; nuisance features shift by 0.6; a new scanner adds benign look-alikes to detector 0; a seven-day campaign raises family 5 prevalence	Prevalence 4.2% before and 5.6% after drift onset	Pendlebury et al. (2019); Jordaney et al. (2017)
Seeds	The seed fixes data, drift, initialisation and sampling; every method sees the same test alerts within a seed	3 seeds	Law (2015); Henderson et al. (2018)

Table 4: Costs, capacity, label delay and the stylised analyst model: definitions and values

Element	Definition	Value or range (stylised)	Basis
Asset criticality tiers	Four tiers by quantile of a latent asset score; loss avoided if a true incident is investigated the same day	USD 100, 400, 1,500, 6,000 for tiers 0 to 3 (shares 38%, 35%, 19%, 8%)	Hubbard & Seiersen (2016); Aven (2016)
Review cost and capacity	Analyst time per alert reviewed; share of daily alerts that can be reviewed	USD 40 per review (loaded cost); 10 minutes of analyst time per review for workload accounting; capacity 12.5% (150 alerts per day); 3% to 25% in sweeps	Alahmadi et al. (2022); Kokulu et al. (2019)
Escalation cost	Cost of a false escalation after review (senior time, disruption)	USD 80 per false escalation (twice the review cost)	Cichonski et al. (2012)
Analyst evidence	Own log-likelihood evidence for incident alerts and benign alerts, with noise	Easy incidents +2.2, hard incidents +0.6, benign −1.0; noise standard deviation 1.5	Assumption; Kahneman (2011)
Reliance on the machine	Weight w on the displayed log-odds, the remainder on the base rate; decision at posterior 0.2	w = 0.7 baseline; raised by 0.15 with explanations; multiplied by 0.5 on flagged alerts	Lee & See (2004); Parasuraman & Manzey (2010); Bansal et al. (2021)
Explanations	Improve own evidence by 25% when top-3 attribution is stable; add review time	Gain 0.25; time 1.2 times review time	Assumption; Poursabzi-Sangdeh et al. (2021)
Uncertainty flag	Mid-range probability (0.15 to 0.85) or ensemble spread in the top third	Flag share and effects reported in Table 13	Bhatt et al. (2021)
Label delay and adaptation	Labels for day d become usable on day d + 2; recalibration on the last five labelled days plus four calibration days	Two-day delay; adaptive conformal step 0.05	Gibbs & Candès (2021)

Detectors, calibrators and uncertainty

Five detectors (Table 5) span the usual choices: logistic regression, a deliberately flexible boosted-tree model that overfits the rare class (one member and a five-member bagged ensemble), a regularised shallow boosted tree, and an ensemble of five small neural networks (Ke et al., 2017; Friedman, 2001; Hastie et al., 2009; Lakshminarayanan et al., 2017). The flexible tree is the main object of study, because it ranks well and is poorly calibrated, which is the situation in which calibration matters. Calibrators are fitted on the calibration window and applied to the test period: temperature scaling, Platt scaling, beta calibration, isotonic regression and histogram binning with 15 equal-mass bins (Guo et al., 2017; Platt, 2000; Kull et al., 2017; Zadrozny & Elkan, 2002). A per-family variant fits one Platt model per alert family, and an adaptive variant refits Platt scaling each test day on the labelled alerts of days d − 7 to d − 2 together with the last four days of the calibration window. A calibration-set size sweep refits each calibrator on the most recent 500 to 12,000 calibration alerts. Calibration is reported as ECE with 15 equal-mass bins, the Brier score and log loss, and by reliability curves over deciles (Naeini et al., 2015; Brier, 1950; Gneiting & Raftery, 2007). Because binned ECE is biased and depends on binning, ECE is read alongside the proper scores and the reliability curves (Nixon et al., 2019; Vaicenavicius et al., 2019). Two uncertainty signals are used. The first is the spread (standard deviation) of the predicted probabilities of the five bagged deep trees, an epistemic proxy. The second is $p(1 - p)$ of the calibrated probability, an aleatoric proxy. Conformal sets use Mondrian (class-conditional) split conformal prediction on the isotonic-calibrated probability: half of the calibration window fits the isotonic map, the other half supplies the quantile of the nonconformity score $1 - p$ of the true class, separately for incidents and benign alerts, at $\alpha = 0.05$, 0.10 and 0.20. A set can be {benign}, {malicious}, {both} or empty; {both} is the ambiguous outcome used as an escalation signal (Vovk et al., 2005; Angelopoulos & Bates, 2023; Papadopoulos et al., 2002). An adaptive variant lowers or raises the working miscoverage level for each class by a step of 0.05 in proportion to the error observed on labels two days old (Gibbs & Candès, 2021). Selective prediction keeps the alerts with the lowest uncertainty and reports the error rate of the kept alerts at a threshold of 0.5 as coverage falls (Chow, 1970; Geifman & El-Yaniv, 2017).

Table 5: Detectors, calibrators, conformal sets, explanation methods and evaluation: specification and purpose

Component	Specification	Purpose	Reference
Logistic regression	L2 (C = 1), standardised inputs	Naturally calibrated baseline	Hastie et al. (2009)
Boosted trees, deep (1 and 5 bagged)	400 trees, learning rate 0.1, 63 leaves, minimum 5 samples per leaf, subsample 0.8, column sample 0.8	Strong ranker that overfits a rare class and gives extreme scores	Ke et al. (2017); Friedman (2001)
Boosted trees, shallow (1)	150 trees, learning rate 0.05, 15 leaves, minimum 40 samples per leaf	Regularised comparison	Ke et al. (2017)
MLP ensemble (5)	Two hidden layers (64, 32), 60 epochs, standardised inputs, mean of five members	Neural comparison and ensemble spread	Lakshminarayanan et al. (2017)
Calibrators	Temperature, Platt, beta, isotonic regression and 15-bin histogram binning on the calibration days; per-family Platt; sliding-window Platt	Post-hoc probability correction	Guo et al. (2017); Platt (2000); Kull et al. (2017); Zadrozny & Elkan (2002)
Conformal sets	Mondrian (class-conditional) split conformal on isotonic-calibrated scores; half of the calibration days fit the calibrator, half set the quantile; $\alpha = 0.05$, 0.10, 0.20	Escalation signal with coverage guarantee under exchangeability	Vovk et al. (2005); Angelopoulos & Bates (2023)
Explanations	TreeSHAP for each of five bagged members; 1,000 test alerts (400 incidents, 600 benign) per seed; top-3 stability by Jaccard index across member pairs	Attribution with spread and stability	Lundberg et al. (2020)
Evaluation	ECE with 15 equal-mass bins, Brier score, log loss, AUROC, AUPRC; queue metrics on 20 test days; 3 seeds	Verification and comparison	Naeini et al. (2015); Brier (1950); Davis & Goadrich (2006)

Prioritisation and economics

The queue reviews, each day, the 150 alerts (12.5%) with the highest priority score. Ten orderings are compared: random; the raw score; the calibrated probability; the raw score times asset loss; the calibrated probability times asset loss; the same with per-family and adaptive calibration; the calibrated probability plus or minus one ensemble spread, times loss (an upper and a lower confidence variant); and an oracle that knows the labels. Ordering by expected loss avoided per unit of review time is the weighted-shortest-processing-time rule, optimal for a single resource when review time is constant (Smith, 1956; Pinedo, 2016). A capacity sweep from 3% to 25% shows how the ordering matters as the constraint tightens. A second experiment drops the capacity limit and applies the cost-optimal threshold rule, review if p × loss exceeds the review cost, with each probability source, including a prior-shift correction using the test period's prevalence odds, which is an oracle quantity used only to bound what correction could achieve (Elkan, 2001; Saerens et al., 2002). Net value is the loss avoided on confirmed incidents less the cost of the reviews and of false escalations. All values are in USD per day of the testbed.

Explanations

Shapley attributions use TreeSHAP for each of the five bagged members on a sample of 1,000 test alerts per seed (400 incidents and 600 benign alerts), with the mean over members as the explanation and the spread as its uncertainty (Lundberg et al., 2020; Lundberg & Lee, 2017; Štrumbelj & Kononenko, 2014). Four properties are measured. Stability is the mean Jaccard index of the top-three features between member pairs. Agreement with ground truth is the share of the top-three features, by mean absolute attribution, that belong to the four truly informative features of the alert's family (its own detector, user risk, peer deviation and indicator hits), which a real SOC cannot know. Faithfulness is the deletion ratio: the mean absolute change in the ensemble logit when the three top-attributed features are replaced by their training means, divided by the same change for three random features; a ratio of one means the attribution is no better than chance (Yeh et al., 2019; Hooker et al., 2019; Bhatt et al., 2020). Surrogate fidelity is the R-squared and AUROC of a depth-three regression tree fitted to the calibrated member probability on calibration alerts and evaluated on the test period (Bastani et al., 2017; Rudin, 2019). Each property is reported by tercile of ensemble spread, by easy and hard incidents, and before and after drift.

Analyst model

No analysts were studied, so the human side is a model whose assumptions are explicit and varied. For each reviewed alert the analyst forms own evidence as a log-likelihood ratio: +2.2 for an easy incident, +0.6 for a hard incident and −1.0 for a benign alert, with Gaussian noise of standard deviation 1.5. The decision combines this with the machine's displayed log-odds, weighted by a reliance weight w (0.7 at baseline) and shrunk towards the base rate otherwise: posterior log-odds equal own evidence plus (1 − w) times the base-rate log-odds plus w times the displayed log-odds, and the analyst escalates when the posterior probability exceeds 0.2 (Lee & See, 2004; Parasuraman & Manzey, 2010). Conditions differ in what is displayed and in the queue. An uncertainty flag (mid-range probability or high spread) halves w for flagged alerts. An explanation raises w by 0.15 and raises own evidence by 25% when the top-three attribution is stable, and it costs 20% more review time; both effects are assumptions motivated by the finding that explanations raise acceptance of advice and can help only when they are informative (Bansal et al., 2021; Zhang et al., 2020; Buçinca et al., 2021). Nine conditions are run on the same test alerts and the same noise draws: the analyst alone with a random queue and with the value queue; the raw score shown with the raw queue and with the value queue; the calibrated probability shown; the calibrated probability with a flag, with an explanation, and with both; and the raw score with an explanation. Reliance is measured as the share of machine advice (the upper half of each day's queue by displayed probability) that the analyst follows when the advice is wrong and when it is right.

Monte Carlo, Sobol indices and metamodel

The decision value is propagated through 2,000 draws of ten inputs sampled uniformly: loss scale (0.3 to 3), review cost (USD 20 to 80), capacity (6% to 20%), attack-rate multiplier (0.5 to 3, applied to incident contributions), drift weight (0 to 1, the weight on post-drift days), baseline reliance w (0.4 to 0.95), explanation gain (0 to 0.4), explanation trust (0 to 0.3), flag shrinkage (0 to 0.8) and explanation time

multiplier (1.0 to 1.5, applied to review and false-escalation costs). Six arms are compared with the baseline of a raw-score-times-loss queue with the raw score displayed: the calibrated queue and display, with a flag, with an explanation, with both, and the calibrated queue with a raw display and an explanation. First-order and total Sobol indices of the gain of the flag-and-explanation arm are estimated with a Saltelli sample of 512 base draws (11,264 evaluations) on the first seed's data (Sobol', 2001; Saltelli et al., 2010; Herman & Usher, 2017). A gradient-boosted regression and a depth-three tree on the 2,000 draws serve as metamodels, with five-fold cross-validated R-squared for the former.

Verification

Checks were run before the experiments were interpreted. The random queue found 12.7% of incidents at a capacity of 12.5%, as it should; the oracle queue found all of them and gave the highest net value; the logistic regression, which is calibrated by construction on this data, had an ECE of 0.005 and a mean probability equal to the prevalence; the drift raised the test prevalence from 4.2% to 5.6% and lowered the AUPRC of every tree and neural model, and Figure 4(c) shows the intended detector shift; conformal sets reached at least the nominal coverage on the calibration half; and the two analyst-alone conditions, whose decisions ignore the display, confirmed the same share of incidents under a random queue and under the value queue (53.4% and 52.8%). One error found during verification is worth recording: an early version computed the benign nonconformity score as $1 - (1 - p)$, which breaks ties through floating-point rounding and made coverage depend on isotonic plateaus; the score is now computed directly, and all results here use the corrected code.

Results

All results come from the synthetic testbed and are means with standard deviations over three seeds unless stated. Absolute values (ECE, recall, USD) describe the testbed and should not be read as estimates for any real operation. The orderings and mechanisms are the findings, and even those hold only under the assumptions in Tables 3 and 4.

Detector quality and raw calibration

Table 6 shows ranking quality and calibration for the five detectors. The shallow boosted tree ranks best (AUROC 0.802 and AUPRC 0.237 before drift), the bagged deep trees next (0.779 and 0.213), and the logistic regression and MLP ensemble lowest (0.765 and 0.758). Ranking quality is modest throughout, because 30% of incidents are hard by design and part of the signal is irreducible. Calibration does not follow ranking. The logistic regression (ECE 0.005) and the shallow tree (0.007) are close to calibrated before drift. The MLP ensemble is moderately off (0.021). The deep trees are the worst (0.034 for a single tree and for the bagged ensemble) and are wrong in a particular direction: their mean predicted probability is 0.84% against a true incident rate of 4.21%, so they under-predict by a factor of five, and the error grows after drift to 0.049 and 0.048. Raw scores of the deep trees therefore rank well and mislead as probabilities, which is the case in which calibration can matter. Figure 5(a) and (b) plot the two measures by detector and drift period.

Table 6: Detectors: ranking quality, mean predicted probability and calibration error before and after drift (means over three seeds)

Model	AUROC pre	AUPRC pre	Mean p (%) pre	ECE pre	ECE post	AUPRC post	Mean p (%) post
Logistic regression	0.765	0.154	4.21	0.005 ± 0.001	0.012 ± 0.003	0.158	4.44
Boosted trees, deep (1)	0.761	0.185	0.84	0.034 ± 0.002	0.049 ± 0.002	0.165	0.76
Boosted trees, deep (5 bagged)	0.779	0.213	0.85	0.034 ± 0.001	0.048 ± 0.002	0.188	0.78
Boosted trees, shallow (1)	0.802	0.237	4.16	0.007 ± 0.001	0.014 ± 0.003	0.215	4.31
MLP ensemble (5)	0.758	0.185	4.37	0.021 ± 0.002	0.028 ± 0.002	0.173	4.29
True incident rate (%)			4.21				5.61

Figure 5: Models and calibrators: (a) ranking quality by detector and drift period, (b) raw calibration error, (c) effect of calibrators on the deep boosted tree, (d) calibration error against the number of calibration alerts

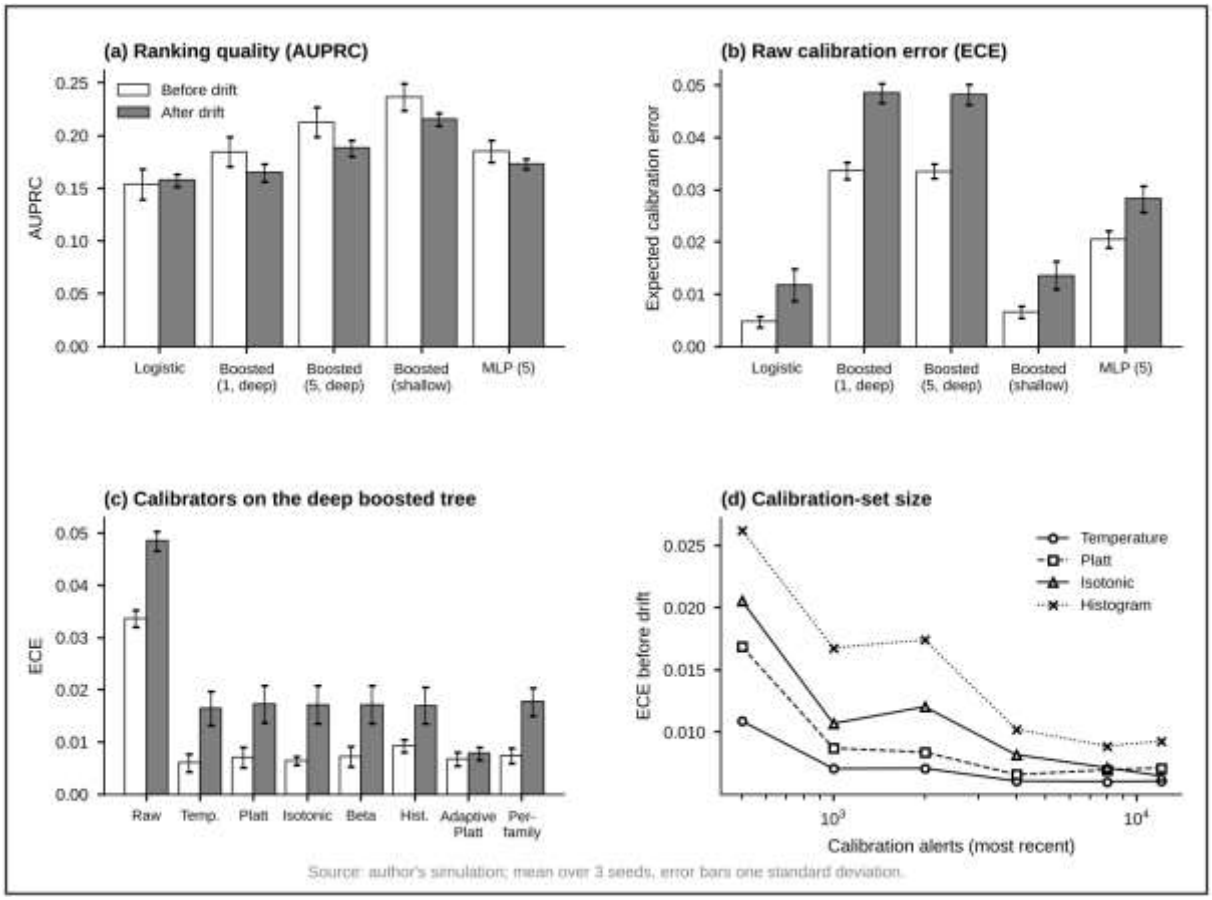


Table 7 and Figure 5(c) give the effect of each calibrator on the deep boosted tree. Every post-hoc method removes most of the error before drift: ECE falls from 0.034 to between 0.006 (temperature scaling) and 0.009 (histogram binning), a reduction of 72% to 82%, and log loss falls from 0.218 to 0.153, a reduction of 30%. The methods are hard to separate at this sample size; temperature, Platt, beta and isotonic differ by less than one standard deviation. Brier scores improve little (0.0394 to 0.0374), because the Brier score is dominated by the rare-event variance that calibration cannot change. AUROC is unchanged by the monotone parametric calibrators and falls slightly with isotonic regression (0.759) and histogram binning (0.755), whose plateaus create ties. After drift every static calibrator degrades to an ECE of about 0.017, so the benefit shrinks from about 82% to about 66% of the raw error. Adaptive Platt scaling, refitted daily on labels two days old, holds ECE at 0.0078, which is 55% below static Platt scaling and 84% below the raw score. Figure 5(d) shows the cost of having few labels. With the most recent 500 calibration alerts, which contain about 20 incidents, temperature scaling has an ECE of 0.011, Platt scaling 0.017, isotonic regression 0.021 and histogram binning 0.026. At 1,000 alerts the values are 0.007, 0.009, 0.011 and 0.017. The non-parametric methods need roughly 4,000 alerts or more to approach the parametric ones (0.006, 0.007, 0.008 and 0.010), and at 12,000 alerts all four lie between 0.006 and 0.009. A one-parameter method is therefore the safer choice when labelled alerts are scarce, which in a SOC is the usual case for rare families. Figure 6(a) and (b) show the reliability curves of the deep tree before and after drift, and Figure 6(c) shows the MLP ensemble, whose mean probability (4.37%) is close to the base rate, so that its error lies in the shape of the curve and not its level.

A per-family Platt model does not improve the global ECE (0.0074 before and 0.0177 after drift), but it matters inside families, as Table 8 shows.

Table 7: Calibrators applied to the deep boosted tree: calibration error, proper scores and ranking quality before and after drift (means over three seeds)

Calibrator (deep boosted tree)	ECE pre	ECE post	Brier pre	Brier post	Log loss pre	AUROC pre
Raw score	0.0337 ± 0.0016	0.0485 ± 0.0019	0.0394	0.0539	0.218	0.761
Temperature scaling	0.0060 ± 0.0017	0.0165 ± 0.0032	0.0374	0.0506	0.153	0.761
Platt scaling	0.0071 ± 0.0020	0.0173 ± 0.0036	0.0374	0.0506	0.153	0.761
Beta calibration	0.0072 ± 0.0019	0.0172 ± 0.0036	0.0374	0.0506	0.153	0.761
Isotonic regression	0.0065 ± 0.0008	0.0171 ± 0.0036	0.0374	0.0506	0.153	0.759
Histogram binning (15)	0.0093 ± 0.0012	0.0170 ± 0.0035	0.0381	0.0509	0.156	0.755
Adaptive Platt (delayed labels)	0.0067 ± 0.0013	0.0078 ± 0.0013	0.0374	0.0504	0.153	0.761
Per-family Platt	0.0074 ± 0.0015	0.0177 ± 0.0027	0.0372	0.0506	0.152	0.766

Figure 6: Reliability curves for the deep boosted tree before drift (a) and after drift (b), for the MLP ensemble before drift (c), and calibration error by alert family (d)

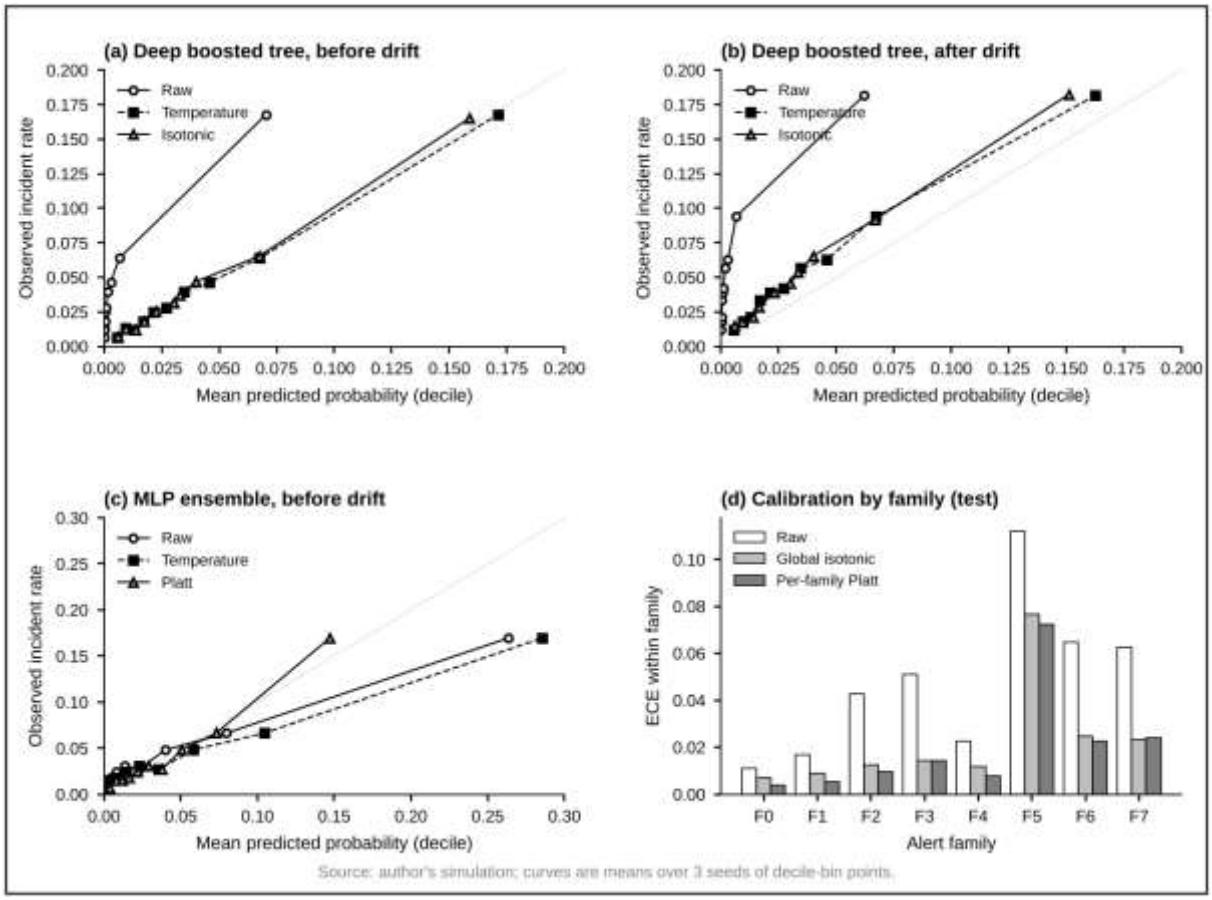


Table 8 and Figure 6(d) give calibration within families. A global calibrator leaves family-level errors of 0.007 to 0.025 in seven families and 0.077 in family F5, the family hit by the seven-day campaign. Per-family Platt scaling lowers the error in the low-volume families F0, F1 and F4 (from 0.007, 0.009 and 0.012 to 0.004, 0.005 and 0.008) and gives a small gain in F5 (0.072), but it leaves F3, F6 and F7 unchanged within noise. Family-level ECE is itself biased upward when a family has few incidents (F0 has 64 on average), so differences of a few thousandths should not be over-read. The main message of Table 8 is that a global calibrator can hide a family that is badly calibrated, and that a surge in a single family is the hardest case for either approach.

Table 8: Calibration error within alert families: raw deep tree, global isotonic regression and per-family Platt scaling (test period, means over three seeds)

Family	Test alerts (mean)	Incidents (mean)	ECE raw	ECE global isotonic	ECE per-family Platt
F0	5,076	64	0.011	0.007	0.004
F1	4,150	80	0.017	0.009	0.005
F2	3,231	163	0.043	0.012	0.010
F3	2,727	170	0.051	0.014	0.014
F4	2,300	60	0.023	0.012	0.008
F5	3,102	368	0.112	0.077	0.072
F6	2,036	199	0.065	0.025	0.023
F7	1,377	108	0.063	0.023	0.024

Conformal sets and uncertainty signals

Table 9 gives the Mondrian conformal results on isotonic-calibrated scores. Class-conditional coverage was at or above its nominal level before drift: 0.969 and 0.980 at $\alpha = 0.05$ (target 0.95), 0.931 and 0.928 at $\alpha = 0.10$ (target 0.90) and 0.835 for both classes at $\alpha = 0.20$ (target 0.80). The excess over target is 3 to 4 points at the two looser levels. I suspect it comes from the finite-sample correction and the plateaus of the isotonic map, but I did not test that. After drift, coverage of incidents fell by 1.5 points at $\alpha = 0.10$ (0.916) and by 4.6 points at $\alpha = 0.20$ (0.789, below the 0.80 target), while coverage of benign alerts barely moved. The adaptive update left incident coverage at 0.910 and brought benign coverage to 0.908 at $\alpha = 0.10$, so both classes sat within about a point of target; but static sets were already close, and the adaptive variant gave more single-label sets that were less accurate (77.0% against 80.1% after drift). Daily coverage of incidents, which rests on about 60 incidents a day, ranged from 0.879 to 0.958 for the static method, so day-to-day variation is as large as the drift effect (Figure 7(a)).

Table 9: Conformal sets: class-conditional coverage, composition of the sets and accuracy of single-label sets, before and after drift (means over three seeds)

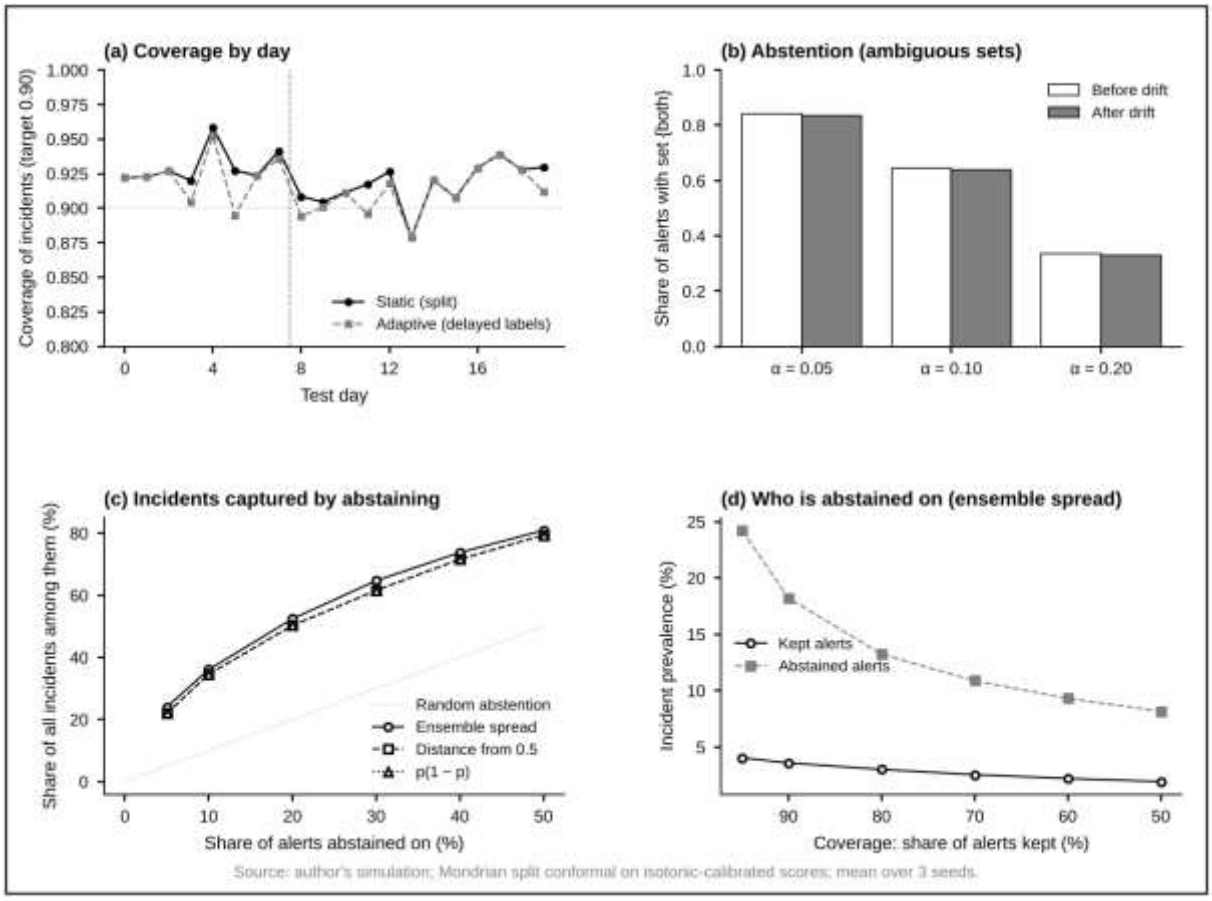
Method	Drift	Coverage, incidents	Coverage, benign	Set {both} (%)	Set {malicious} (%)	Set {benign} (%)	Accuracy of single-label sets (%)
Static, $\alpha = 0.05$	Before	0.969	0.980	84.1	2.7	13.2	81.4
	After	0.966	0.982	83.5	2.4	14.1	83.0
Static, $\alpha = 0.10$	Before	0.931	0.928	64.5	8.5	27.0	79.7
	After	0.916	0.929	63.8	8.3	27.9	80.1
Static, $\alpha = 0.20$	Before	0.835	0.835	33.5	18.0	48.4	75.0
	After	0.789	0.835	32.9	18.1	48.9	75.0
Adaptive, $\alpha = 0.10$	Before	0.925	0.922	61.7	9.1	29.2	79.8
	After	0.910	0.908	60.0	10.5	29.5	77.0

The sets are valid but largely uninformative. Because the detector's AUROC is only about 0.76, most alerts cannot be placed in one class with the required confidence: the share of {both} sets is 84% at $\alpha = 0.05$, 64.5% at $\alpha = 0.10$ and 33.5% at $\alpha = 0.20$. Single-label sets are right about 75% to 81% of the time, which is below $1 - \alpha$, as it should be: the guarantee is a class-conditional coverage property and not a statement about the accuracy of single-label sets. An ambiguous set is therefore a weak escalation signal in this testbed, since it applies to roughly two-thirds of alerts at the 10% level (Figure 7(b)).

The ensemble spread behaves more as a risk indicator than as an epistemic alarm. Its mean is 0.035 for easy incidents, 0.0096 for hard incidents and 0.0049 for benign alerts, so it rises with the score level, not

with the difficulty of the case, and a hard incident looks only twice as uncertain as a benign alert. Its rank correlation with the absolute error of the calibrated probability is 0.81, but because 95% of alerts are benign and their absolute error equals their probability, this mostly reflects the fact that spread rises with the score. After drift the mean spread was 0.965 of its pre-drift value, so it did not rise when the data shifted. As a basis for abstention, however, the spread is slightly better than the probability alone (Figure 7(c)): abstaining on the 5% of alerts with the highest spread captures 24% of all incidents (their prevalence is 24%, against 5.05% overall), on 10% captures 36% and on 20% captures 52%, against 22%, 35% and 50% for abstention by probability and 5%, 10% and 20% for random abstention. The gain of about two points comes from an ordering that is close to the probability's, and the abstained set is where the incidents are, as Figure 7(d) shows. The error rate at the 0.5 threshold is, in this testbed, essentially the prevalence among the kept alerts, since calibrated probabilities rarely exceed 0.5: selective prediction here is a form of risk-based escalation and not a classifier with a reject option in the classical sense (Chow, 1970; Geifman & El-Yaniv, 2017).

Figure 7: Conformal sets and uncertainty signals: (a) daily coverage of incidents, static and adaptive, (b) share of ambiguous sets by miscoverage level, (c) share of all incidents captured by abstaining on the highest-uncertainty alerts, (d) incident prevalence among kept and abstained alerts



Prioritisation under capacity

Table 10 and Figure 8 compare the queue orderings at a capacity of 12.5% (150 reviews a day). Ordering by the calibrated probability gives the same result as ordering by the raw score: 39.4% against 39.2% of incidents found and a net value of USD 16.0 thousand against 15.8 thousand a day, differences far inside one standard deviation. This is expected, since a monotone transformation of the score does not change a ranking. The random queue finds 12.7% of incidents and earns 1.4 thousand; the oracle finds all and earns 53.3 thousand.

The picture changes once the score is multiplied by the asset loss. Ordering by the raw score times loss

earns 29.4 thousand a day and captures 59.7% of the total loss, and ordering by the calibrated probability times loss earns 34.5 thousand and captures 68.3%, a gain of 5.1 thousand a day (17%). The share of the highest-tier incidents that are missed falls from 27.6% to 6.9%. This queue finds fewer incidents (24.5% against 39.4%) and has lower precision (9.9% against 15.9%), because it deliberately prefers rarer but costlier incidents; counting alerts would call it worse. Per-family calibration raises the net value to 36.0 thousand (4.7% of top-tier incidents missed). Adaptive calibration (34.7) is not distinguishable from the static one in this window. Adding one ensemble spread to the probability gives 35.0 and subtracting it gives 32.7, so a conservative upper bound is no worse than the point estimate and a lower bound is worse by 1.8 thousand. The calibrated value queue reaches 65% of the oracle's net value and the per-family variant 68%. Figure 8(a) shows that the advantage of calibration is present at every capacity from 3% to 25% and is largest at 9% to 12.5% of alerts, where the loss captured is 8.6 points higher than with raw scores times loss (60.8% against 52.1% at 9%, 68.3% against 59.7% at 12.5%); it narrows to 4.6 points at 25%, where a larger share of all alerts is simply reviewed.

Table 10: Queue orderings at a capacity of 12.5% of daily alerts: incidents found, precision, loss captured, top-tier incidents missed and net value (test period, means over three seeds)

Queue ordering (capacity 12.5%)	Incidents found (%)	Precision of queue (%)	Loss captured (%)	Tier-3 incidents missed (%)	Net value (thousand USD/day)
Random order	12.7 ± 0.5	5.2 ± 0.3	12.5 ± 1.1	88.7 ± 2.2	1.4 ± 0.7
Raw score	39.2 ± 0.4	15.9 ± 0.6	36.8 ± 1.1	66.4 ± 2.7	15.8 ± 0.6
Calibrated p	39.4 ± 0.6	15.9 ± 0.7	37.1 ± 0.9	66.1 ± 2.3	16.0 ± 0.5
Raw score × asset loss	33.7 ± 0.2	13.6 ± 0.4	59.7 ± 0.5	27.6 ± 1.7	29.4 ± 0.4
Calibrated p × asset loss	24.5 ± 0.5	9.9 ± 0.3	68.3 ± 1.6	6.9 ± 1.2	34.5 ± 0.6
Per-family calibrated p × loss	25.2 ± 1.1	10.2 ± 0.6	70.9 ± 1.0	4.7 ± 0.8	36.0 ± 0.3
Adaptive calibrated p × loss	23.6 ± 0.7	9.5 ± 0.2	68.6 ± 1.7	5.7 ± 0.9	34.7 ± 0.8
Calibrated p + spread, × loss	26.2 ± 0.9	10.6 ± 0.4	69.2 ± 1.7	6.3 ± 1.6	35.0 ± 0.7
Calibrated p − spread, × loss	22.2 ± 0.8	9.0 ± 0.3	65.3 ± 1.7	9.4 ± 0.6	32.7 ± 0.7
Oracle	100.0 ± 0.0	40.4 ± 1.3	100.0 ± 0.0	0.0 ± 0.0	53.3 ± 0.4

When capacity is not fixed and the cost-optimal rule is used instead (review if p × loss exceeds USD 40), calibration changes the decision itself (Table 11). The raw score of the deep tree triggers 34 reviews a day, finds 16.1% of incidents and yields a net value of 16.5 thousand; any calibrated probability triggers 230 to 249 reviews a day, finds 39% to 40% and yields 31.4 to 32.0 thousand, roughly twice the value. After drift the raw score finds 10.0% and yields 14.7 thousand, and the calibrated sources yield 41.5 to 43.1 thousand (temperature 41.5, isotonic 42.1, per-family 43.1, adaptive 42.3, with the adaptive rule reviewing 285 alerts a day against 231 to 247 for the static ones). A prior-shift correction using the true test prevalence, which is an oracle quantity, raises recall to 40.8% with 315 reviews a day but not the net value (41.9 against 42.1), a sign that the rule is flat near its optimum. The extra reviews are worth about what they cost.

Figure 8: Triage: (a) loss captured against review capacity, (b) top-tier incidents missed, (c) net value by queue ordering, (d) net value of the cost-optimal review rule after drift by source of probability

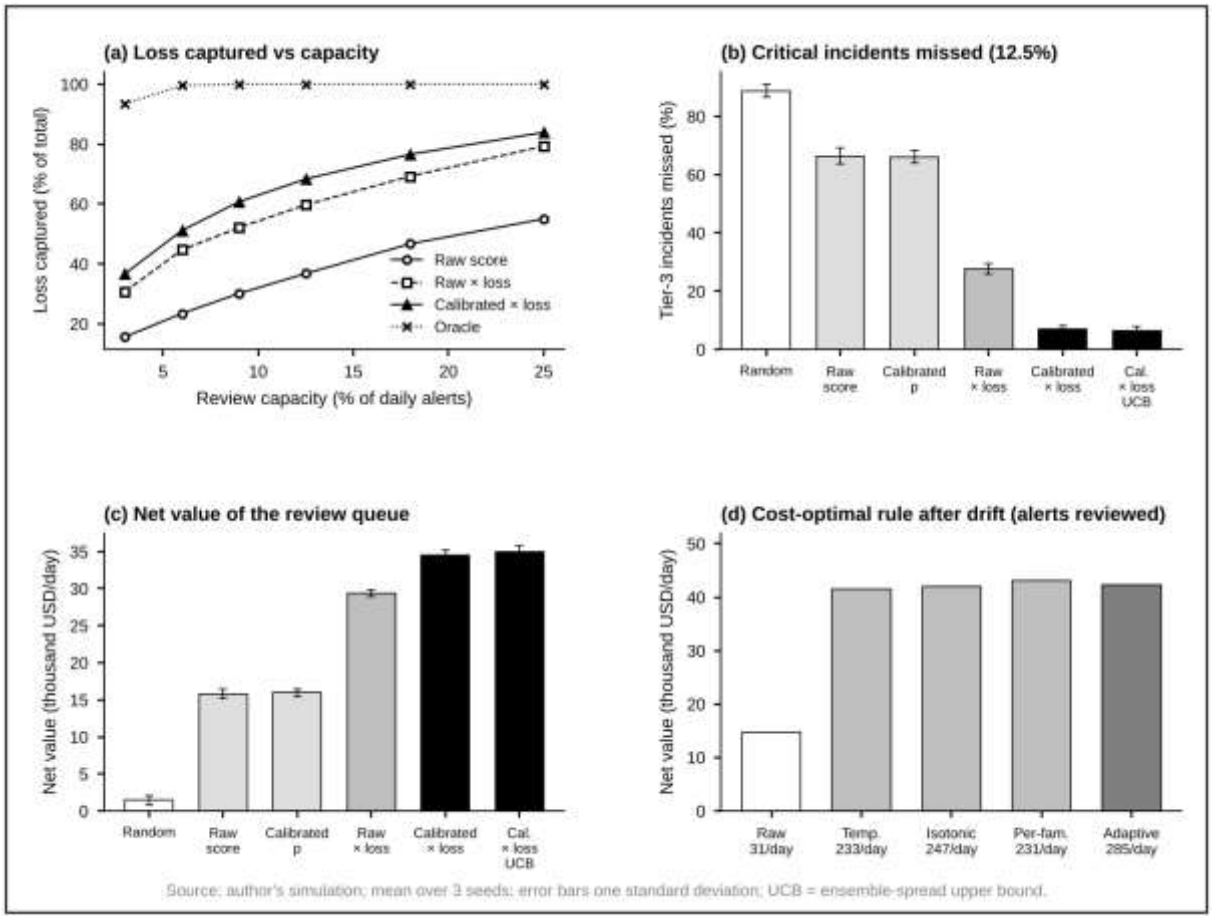


Table 11: Cost-optimal review rule by source of probability: reviews per day, incidents found and net value before and after drift (means over three seeds)

Probability used (rule: review if p × loss > USD 40)	Reviews/day pre	Incidents found (%) pre	Net (k USD/day) pre	Reviews/day post	Incidents found (%) post	Net (k USD/day) post
Raw score	34	16.1	16.5	31	10.0	14.7
Temperature	236	40.1	32.0	233	33.1	41.5
Isotonic	249	39.3	31.4	247	33.2	42.1
Per-family Platt	230	39.0	31.9	231	32.8	43.1
Adaptive Platt	242	39.4	31.7	285	36.4	42.3
Isotonic with prior-shift correction	n/a	n/a	n/a	315	40.8	41.9

Explanations

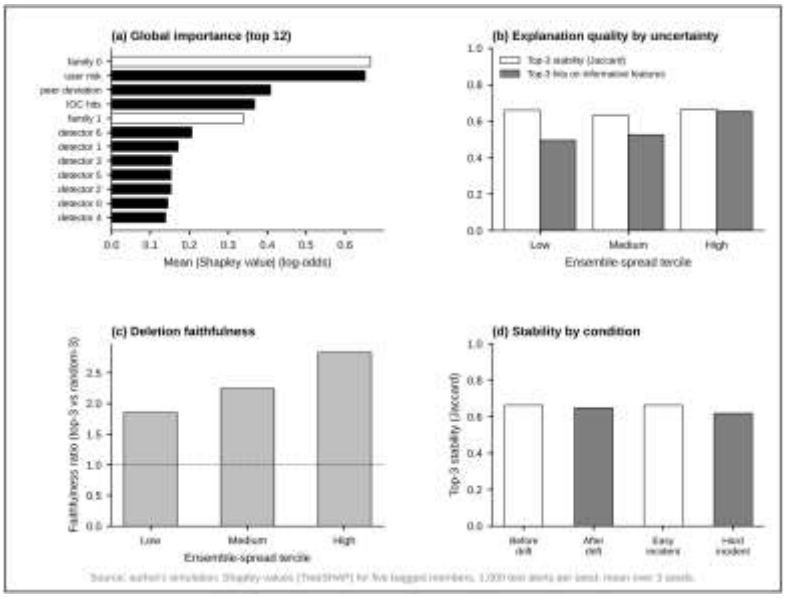
Table 12 summarises the Shapley explanations of the five bagged members. Top-three attributions agree across members with a mean Jaccard index of 0.654 ± 0.018, which is moderate: typically two or three of the three leading features coincide. The index is similar for incidents and benign alerts (0.65 and 0.66), a little lower for hard incidents (0.62 against 0.66 for easy ones) and the same before and after drift (0.66 and 0.65). It is unrelated to the ensemble spread (Spearman 0.027 ± 0.061), so the spread cannot be used to flag explanations that are unstable. On incidents, 59% of the top-three features belong to the truly informative features of the alert's family, 62% for easy and 52% for hard incidents, and the

share rises with spread (0.49, 0.52 and 0.65 across terciles). Deletion faithfulness is 2.39 ± 0.29: removing the top-three features changes the ensemble logit about 2.4 times as much as removing three random ones, and the ratio also rises with spread (1.85, 2.24 and 2.83). These increases are not evidence that uncertain alerts are better explained. High-spread incidents are mostly easy incidents with high scores and strong signals, so spread tercile is confounded with difficulty; I could not separate the two with the present design. Attributions place 15.8% of their mass on nuisance features, against 25% if all features were used equally, which suggests that the explanations find signal better than chance and still pick up some of the overfitting of the deep trees. A depth-three surrogate tree has an R-squared of 0.23 and an AUROC of 0.69 against the 0.76 to 0.78 of the ensemble, so a rule short enough for an analyst to read recovers only a small part of the model's behaviour (Bastani et al., 2017; Rudin, 2019).

Table 12: Explanation diagnostics: stability, agreement with ground truth, deletion faithfulness, attribution spread, nuisance share and surrogate fidelity (means over three seeds)

Diagnostic	Mean ± SD over seeds	Detail
Top-3 stability across five members (Jaccard)	0.654 ± 0.018	Incidents 0.65; benign 0.66; easy 0.66; hard 0.62; before drift 0.66; after 0.65
Correlation of stability with ensemble spread (Spearman)	0.027 ± 0.061	No monotone link: spread does not flag unstable explanations
Top-3 hits on truly informative features (incidents)	0.590 ± 0.028	Easy 0.62; hard 0.52; by spread tercile 0.49, 0.52, 0.65
Deletion faithfulness ratio (top-3 versus random-3 removal)	2.39 ± 0.29	By spread tercile 1.85, 2.24, 2.83
Attribution spread relative to magnitude (top-3, mean of SD/|mean|)	0.161 ± 0.007	Member-to-member variation is moderate for the leading features
Share of attribution on nuisance features	0.158 ± 0.005	Eight of 32 features; uniform use would give 0.25
Depth-3 surrogate: R^2 against calibrated probability; AUROC	0.23; 0.69	A three-level rule recovers a small part of the model

Figure 9: Explanations: (a) global importance of the twelve leading features, (b) stability and agreement with informative features by ensemble-spread tercile, (c) deletion faithfulness by tercile, (d) stability before and after drift and for easy and hard incidents



Analyst conditions

Table 13 and Figure 10 give the results of the nine analyst conditions. Every statement in this section is a consequence of the assumptions in Table 4 and is a hypothesis about people and not a finding about them. Confirmed incidents are the share of incidents in the queue that the analyst escalates. The analyst alone confirms 53% of the incidents in the queue, whether the queue is random or ordered by value, as it must, since own evidence is the same. The queue changes the money: the random queue loses 2.2 thousand a day and the value queue earns 15.2 thousand.

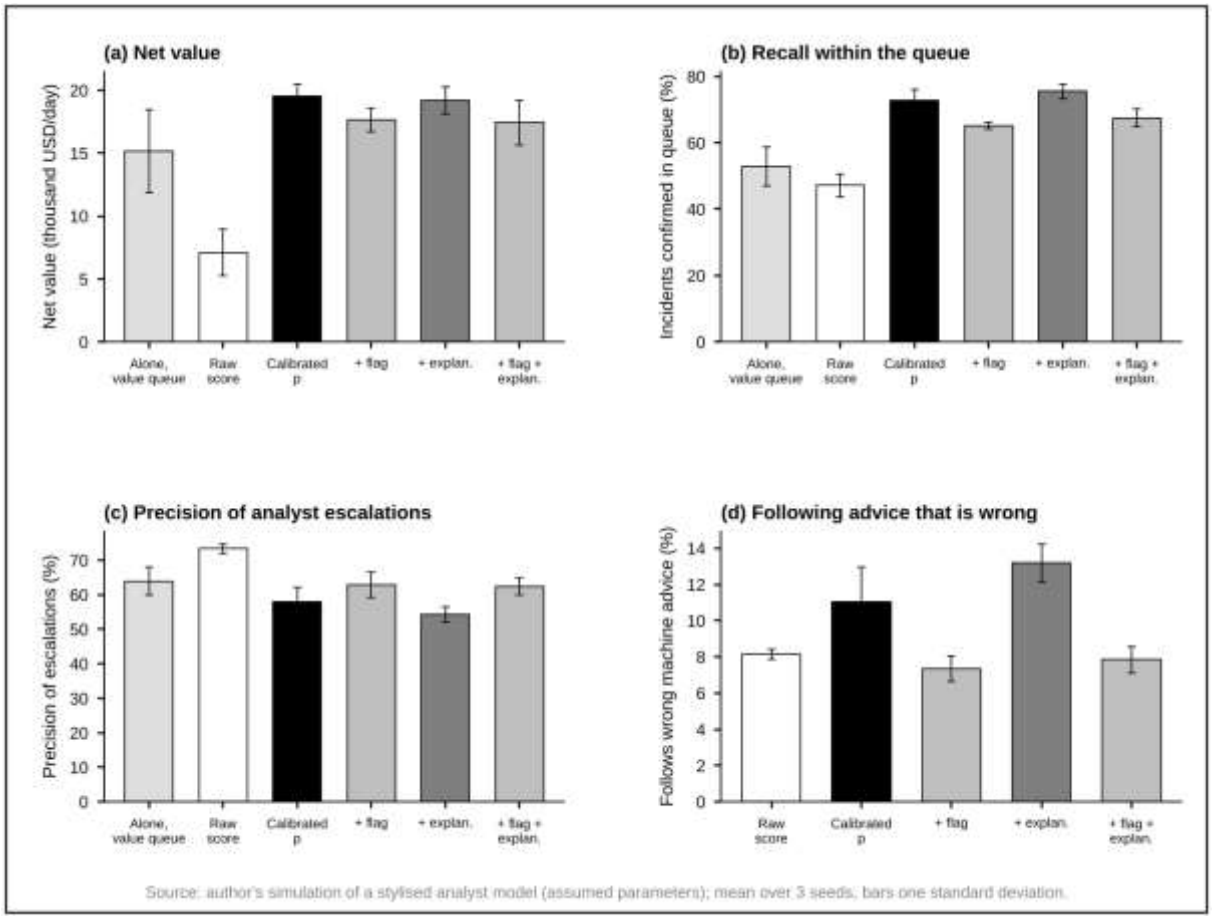
Showing the raw score hurts. With the value queue, the raw score lowers net value to 7.1 thousand, 8.1 thousand below the analyst alone, and confirms only 47.2% of the incidents in the queue. In the model the raw score is about five times too low, and a display that is too low pulls the posterior of ambiguous alerts towards benign. Showing the calibrated probability raises confirmed incidents to 72.9% and net value to 19.5 thousand, which is 12.4 thousand above the raw display and 4.3 thousand above the analyst alone (whose standard deviation over seeds is 3.3 thousand). With the calibrated display the analyst follows wrong advice in 11.1% of cases and right advice in 94.9%.

An uncertainty flag shifts the balance from recall to precision: confirmed incidents fall to 65.0%, precision rises from 57.9% to 62.9%, following wrong advice falls from 11.1% to 7.4%, and net value falls by 1.8 thousand to 17.7 thousand. An explanation does the reverse: confirmed incidents rise to 75.5%, precision falls to 54.3% and following wrong advice rises to 13.2%, because the explanation increases reliance, while the 20% longer review raises analyst time from 25 to 30 hours a day. Its net value (19.2 thousand) is not distinguishable from the calibrated display alone (19.5 thousand). Flag and explanation together land between the two (67.5% confirmed, 62.4% precision, 17.4 thousand). An explanation attached to the raw score does not rescue the raw display (4.9 thousand).

Table 13: Analyst conditions under the stylised analyst model: incidents confirmed, precision, reliance on wrong and right advice, analyst hours and net value (means over three seeds)

Condition (same capacity of 150 reviews/day)	Incidents confirmed (%)	Precision (%)	Follows wrong advice (%)	Follows right advice (%)	Analyst hours/day	Net (k USD/day)
Analyst alone, random order	53.4 ± 2.2	44.5 ± 2.4	3.6 ± 0.3	53.4 ± 2.2	25.0 ± 0.0	−2.2 ± 0.4
Analyst alone, value queue	52.8 ± 5.9	63.9 ± 4.1	3.3 ± 0.3	52.8 ± 5.9	25.0 ± 0.0	15.2 ± 3.3
Raw score queue and display	53.0 ± 2.4	73.2 ± 3.9	12.0 ± 1.1	93.2 ± 0.3	25.0 ± 0.0	8.2 ± 0.3
Value queue, raw score shown	47.2 ± 3.5	73.3 ± 1.3	8.2 ± 0.3	94.0 ± 0.4	25.0 ± 0.0	7.1 ± 1.8
Value queue, calibrated p shown	72.9 ± 3.2	57.9 ± 4.4	11.1 ± 1.9	94.9 ± 0.6	25.0 ± 0.0	19.5 ± 1.0
+ uncertainty flag	65.0 ± 1.1	62.9 ± 3.6	7.4 ± 0.7	92.8 ± 0.1	25.0 ± 0.0	17.7 ± 0.9
+ explanation	75.5 ± 2.1	54.3 ± 2.2	13.2 ± 1.0	95.6 ± 0.1	30.0 ± 0.0	19.2 ± 1.1
+ flag + explanation	67.5 ± 2.6	62.4 ± 2.4	7.9 ± 0.7	93.3 ± 0.4	30.0 ± 0.0	17.4 ± 1.8
Raw score shown + explanation	46.0 ± 5.2	70.0 ± 1.2	8.7 ± 0.2	93.8 ± 1.2	30.0 ± 0.0	4.9 ± 2.8

Figure 10: Analyst conditions under the stylised model: (a) net value, (b) incidents confirmed in the queue, (c) precision of escalations, (d) following wrong machine advice



Monte Carlo and Sobol analysis

Table 14 and Figure 11 propagate the decision value through the ten uncertain inputs. The baseline

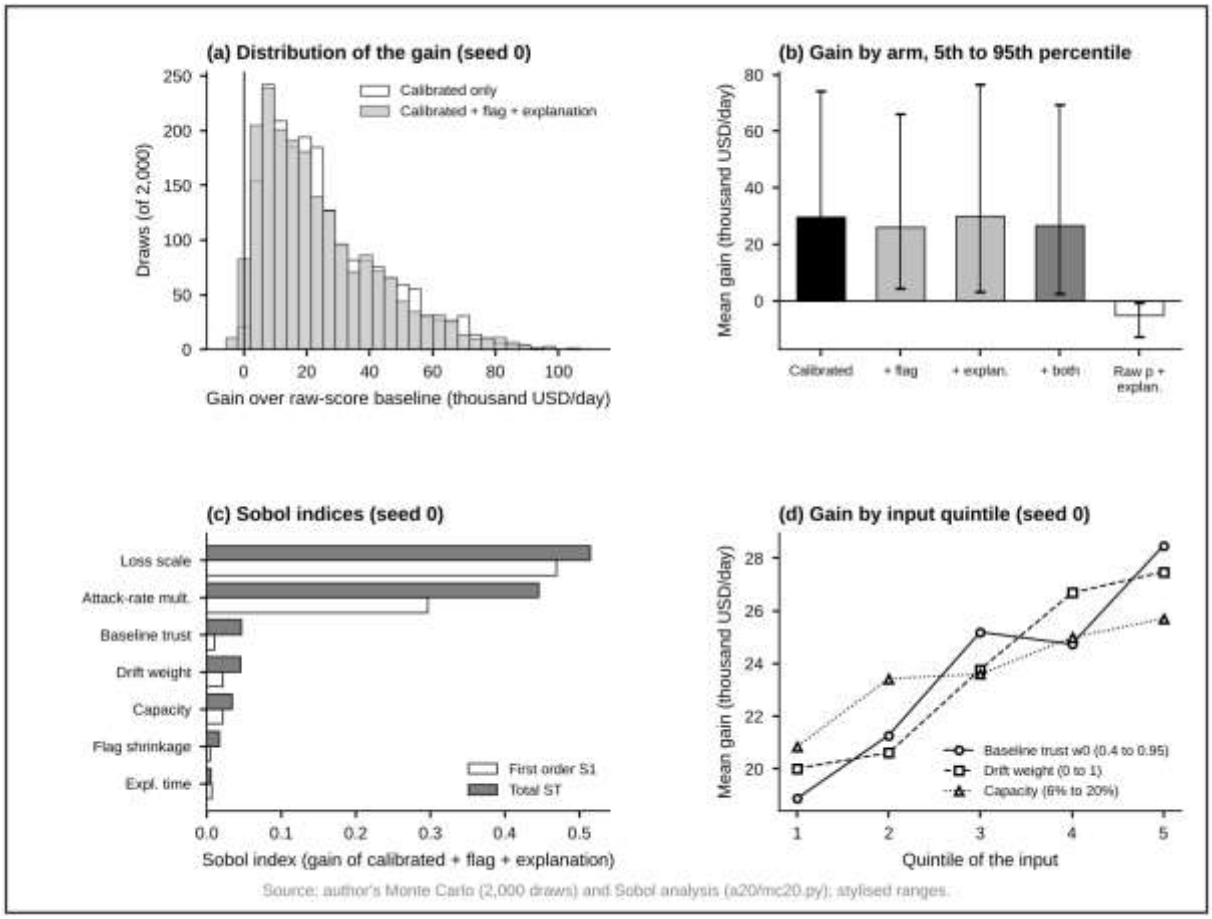
(raw-score queue with raw display) has a mean net value of 33.9 thousand a day over the draws. A calibrated queue and display adds 29.7 thousand on average (5th to 95th percentile 5.0 to 74.3 thousand) and is positive in essentially every draw. Adding a flag lowers the mean gain to 26.1 thousand, an explanation leaves it at 29.9 thousand, and both give 26.4 thousand, so the differences among the calibrated arms are small against the spread across draws. A raw score with an explanation on a calibrated queue loses 5.2 thousand on average and gains in only 3% of draws.

The variance of the gain is mostly a matter of the setting. For the flag-and-explanation arm, loss scale has a first-order index of 0.47 (total 0.52) and the attack-rate multiplier 0.30 (total 0.45); baseline trust, drift weight and capacity each have totals of 0.03 to 0.05. First-order indices sum to 0.84, so about 16% of the variance is interaction, mainly between loss scale and attack rate, which multiply each other. The explanation gain, explanation trust, flag shrinkage and time multiplier together have a total index below 0.03 for each, so the assumed human parameters hardly affect the size of the gain of this arm. A gradient-boosted metamodel reproduces the draws with a cross-validated R-squared of 0.97, and a depth-three tree with 0.75, using loss scale and attack rate for every split. In the first seed the gain rises from USD 18.9 thousand to 28.5 thousand across quintiles of baseline trust, from 20.0 to 27.5 thousand across quintiles of drift weight and from 20.8 to 25.7 thousand across quintiles of capacity, and it is almost flat in review cost (24.9 to 23.1 thousand). A calibrated display is worth more when analysts heed it, when more of the period is post-drift and when more can be reviewed.

Table 14: Monte Carlo and Sobol analysis: gain of each arm over the raw-score baseline, variance shares and metamodel fit (2,000 draws; Sobol on the first seed)

Arm (gain over raw-score baseline)	Mean gain (k USD/day)	5th to 95th percentile	Draws with gain > 0 (%)
Calibrated queue and display	29.7	5.0–74.3	100
+ uncertainty flag	26.1	4.4–66.2	100
+ explanation	29.9	3.3–76.7	99
+ flag + explanation	26.4	2.7–69.3	99
Raw score shown + explanation (calibrated queue)	−5.2	−12.7 to −0.6	3
Baseline (raw-score queue and display): mean net value	33.9		
Sobol indices (gain of flag + explanation arm): S1 / ST	loss scale 0.47 / 0.52; attack-rate multiplier 0.30 / 0.45; baseline trust 0.01 / 0.05; drift weight 0.02 / 0.05; capacity 0.02 / 0.03		
Metamodel: gradient-boosted regression, 5-fold cross-validated R^2 (depth-3 tree R^2)	0.97 (0.75)		

Figure 11: Monte Carlo and Sobol analysis: (a) distribution of the gain, (b) mean gain by arm with 5th to 95th percentiles, (c) Sobol indices for the flag-and-explanation arm, (d) mean gain by quintile of selected inputs



Discussion

Answers to the research questions

RQ1 asked how well alert scores are calibrated and how the repairs hold up under drift, across families and with few labels. The raw scores of the deep trees under-predicted incidents by a factor of five, and any post-hoc calibrator removed 72% to 82% of that error on a window from the same period. Under drift the static calibrators kept about two-thirds of the benefit, daily recalibration with delayed labels kept most of it, and a global calibrator left a family-level error three to eleven times larger in the family hit by the campaign (0.077 against 0.007 to 0.025). With few labels a one-parameter method did best. RQ2 asked what uncertainty signals add. Conformal sets kept class-conditional coverage close to target, but were ambiguous for most alerts; ensemble spread was a modest improvement over probability for choosing alerts to set aside, did not rise with drift, and was unrelated to explanation stability. RQ3 asked about the decision value of calibrated, uncertainty-aware prioritisation under capacity. Calibration alone did not change a ranking. When the probability was multiplied by asset loss, the value of the queue rose 17%, and under a cost-optimal rule the calibrated probability about doubled the net value. Adding a spread bound or per-family and adaptive calibration changed the net value by between −1.8 and +1.5 thousand a day around 34.5 thousand. RQ4 asked about explanations and about displays for analysts. Explanations were moderately stable (0.65) and better than chance on faithfulness (2.4), and their stability was not tied to uncertainty. Under the stylised analyst model, the calibrated display was worth much more than the raw display, whereas flags and explanations moved the balance between recall and precision without raising net value.

Why calibration matters only when probabilities are used as probabilities

The triage result is stated more precisely as follows. A calibrated probability and a raw score give the same queue when the queue is a ranking by probability alone. They differ when the number is

combined with something else: multiplied by an asset loss, compared with a review cost, shown to a person, or averaged across families (Ahmed, 2024; Islam, 2024; Asheke, 2024). A constant factor would not reorder a product, so the effect needs error that varies across the score range, and the raw deep-tree scores are compressed and uneven in that way: an alert on a cheap asset with a high raw score and one on a critical asset with a moderate score trade places differently under raw and true probabilities (Ashrafuzzaman, 2024; Jannatul, 2024; Kynatun & Saba, 2024). Against a fixed review cost even a constant factor matters, since a score five times too small puts most alerts below the threshold, which is the doubling seen under the cost-optimal rule (Elkan, 2001; Vickers & Elkin, 2006; Van Calster et al., 2019). Both mechanisms need the number to be used as a number. I did not run the queue with the shallow tree, whose raw score is already close to calibrated (ECE 0.007), so the inference that a well-calibrated raw score would do as well as a recalibrated one is plausible but untested here (Mashiat, 2024; Sumaiya, 2024; Feroz, 2024). The practical point is that ECE should be reported for the model that will be used with a loss or a cost, and that the value of calibration should be evaluated in that setting, not on the accuracy or AUROC that calibration leaves unchanged (Gneiting & Raftery, 2007; Nixon et al., 2019).

Calibration under drift and with few labels

The drift in this study was one moderate episode, and static calibrators lost about a third of their benefit in it. Adaptive recalibration with labels two days old recovered most of that, which supports a practice of scheduled recalibration with explicit monitoring of ECE and prevalence by week. Two caveats apply. The recalibration labels were complete and correct, whereas real SOC labels exist mainly for alerts that were reviewed, and the reviewed alerts are the highest-ranked ones; a model recalibrated on them will tend to learn the selection and not the truth, which is why a randomised audit sample of unreviewed alerts is among the requirements (Lakkaraju et al., 2017; Ovadia et al., 2019; Gama et al., 2014). And the benefit of any calibration depends on the amount of data: with 500 recent alerts, containing about 20 incidents, only temperature scaling stayed near 0.01. For rare families this suggests pooling alerts across families with a family indicator or a shared temperature, and refining only when sample sizes allow, which is consistent with the modest and uneven gains of per-family models in Table 8 (Kull et al., 2017; Guo et al., 2017).

Uncertainty signals

The conformal sets were valid and uninformative, because the detector was weak. Validity here has two limits. Coverage was class-conditional and marginal over alerts, so it said nothing about coverage for a single family or asset tier, and under drift coverage of incidents fell below target at α = 0.20, as one would expect when a shift breaks exchangeability, while the same shift had only a small effect at α = 0.10 (Tibshirani et al., 2019; Barber et al., 2023; Gibbs & Candès, 2021). The spread gave a risk-ordered abstention that was a little better than the probability, but it did not behave as an epistemic signal: it was highest for easy incidents, lowest for benign alerts, and flat across the drift. This is a caution for practice. Ensemble disagreement is often described as model uncertainty, and in a boosted-tree ensemble trained on the same data it may mostly reflect the score level, so a team that wants a drift alarm should monitor calibration error and feature distributions directly and not rely on the spread (Hüllermeier & Waegeman, 2021; Lakshminarayanan et al., 2017; Ovadia et al., 2019).

Explanations as a control

The explanations in this study were neither useless nor trustworthy by default. They were moderately stable, better than chance at finding the informative features and at changing the output when the attributed features were removed, and a short surrogate tree explained little. The finding that stability did not depend on uncertainty matters for how a display is designed: an uncertainty flag cannot stand in for an explanation-quality flag, so any note on stability must be computed from the members directly, as it was for the display in Figure 2. The deletion test replaces features with training means, and that creates off-distribution inputs, so the faithfulness ratio is an indicator and not a proof (Hooker et al., 2019; Ghorbani et al., 2019; Krishna et al., 2022). The NISTIR 8312 principles of explainable AI separate the accuracy of an explanation from the limits of the system's knowledge, and the results support the separation: faithfulness and stability are properties of the explanation, spread and coverage are properties of the prediction, and they vary independently (Phillips et al., 2021; Adebayo et al., 2018).

Human decision-making

The analyst-model results are hypotheses. They follow from assumed weights, and I make no claim that SOC analysts behave this way. Within those assumptions they show three things. First, a display can do harm: the raw display lowered the net value below the analyst working alone, because its miscalibration was imported into the analyst's decision, in line with the observation that people over-rely on advice that appears confident and under-rely on advice that appears weak (Parasuraman & Manzey, 2010; Lee & See, 2004; Skitka et al., 1999). Second, a calibrated probability avoided that harm (12.4 thousand above the raw display) and added 4.3 thousand over the analyst alone, a gain about the size of that condition's variation over seeds, and it did so without any added review time. Third, flags and explanations traded one error for another: the flag reduced both following of wrong advice and confirmation of true incidents, the explanation did the opposite, and neither raised net value, in line with studies in which explanations increased acceptance of advice whether or not the advice was right (Bansal et al., 2021; Buçinca et al., 2021; Poursabzi-Sangdeh et al., 2021). Whether a real flag lowers reliance only when the advice is wrong depends on how well the flag separates wrong from right advice; in the model it lowered both. Controlled studies with analysts, using real tasks and the measures of reliance in Table 13, are needed before any of these effects are used for design (Green & Chen, 2019; Kocielnik et al., 2019; Alahmadi et al., 2022).

Implications for operators

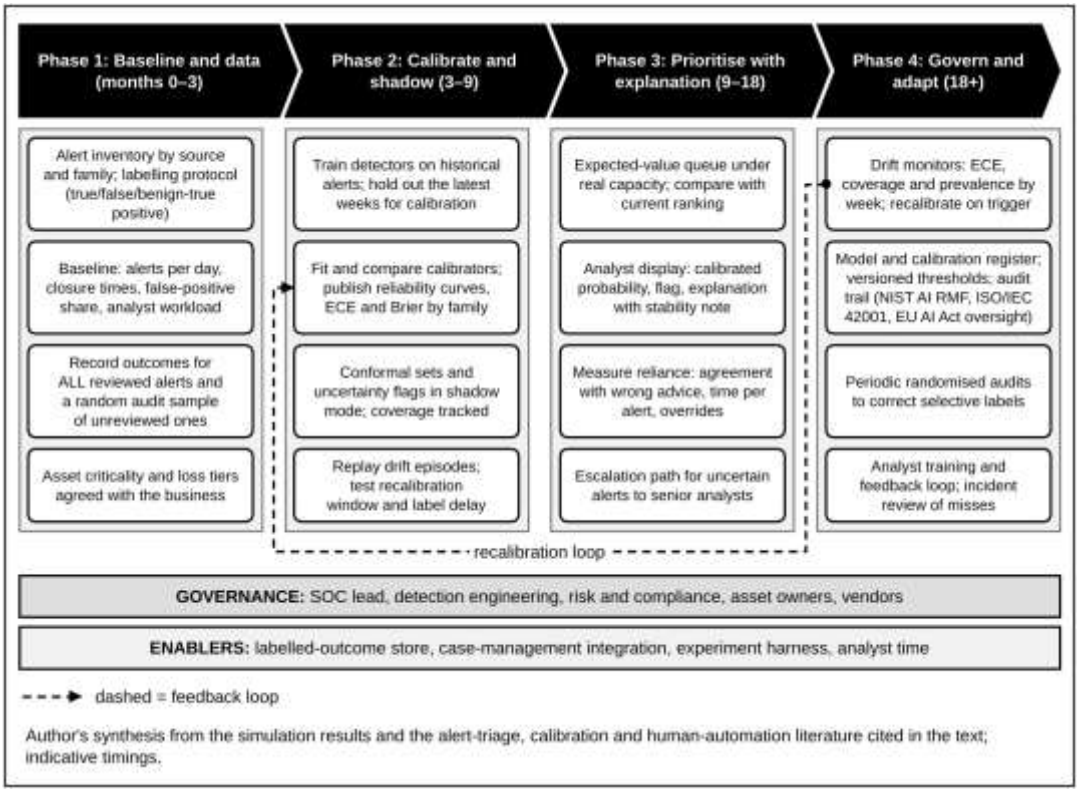
Five implications follow, each with its limit. Measure calibration for the score that feeds a loss or a threshold, and report it by family, since a global figure hid a family in which the error was up to eleven times larger. Rank by expected value, not by probability or count: on this testbed the calibrated value queue found 24.5% of incidents against 39.4% for the probability ranking, yet captured 68.3% of the loss against 37.1%, so teams that report alerts closed or incidents found will mistake it for a regression, and the metric that follows the objective is loss captured or net value at the capacity actually available (Smith, 1956; Pinedo, 2016; Kleinberg et al., 2018). Use the cheapest calibrator that the labelled data supports, refresh it on a schedule, and keep an audit sample of unreviewed alerts so that the labels are not only those of the alerts that were ranked highest. Show a calibrated probability to analysts and test any flag or explanation as a separate intervention, measuring both recall and precision, since each moved them in opposite directions. And treat the base-rate arithmetic as part of the design: at an

incident rate of about 4% the precision of even a good queue is low, as the 9.9% to 15.9% in Table 10 shows, which is why alert fatigue is a property of the situation and not only of the detector (Axelsson, 2000; Cvach, 2012; Sundaramurthy et al., 2015).

Implications for standards, regulation and governance

The results connect to the standards in four ways, and some of those standards have changed or may change, so the citations should be checked against current versions. The incident-handling guide cited here, NIST SP 800-61 Revision 2, treats prioritisation, as I read it, as a function of the functional and informational impact of an incident and the recoverability from it, which is the role asset criticality plays in the value queue (Cichonski et al., 2012; ISO/IEC, 2023b). I believe that NIST has since replaced that revision with a new one aligned to the Cybersecurity Framework 2.0, but I have not verified its current status or content, and a reader applying this work should consult the NIST publication page (National Institute of Standards and Technology, 2024). The NIST AI Risk Management Framework and ISO/IEC 42001 and 23894 ask organisations to measure, document and manage the risks of an AI system, and calibration error, coverage, explanation stability and reliance are measurable properties that fit that record; the governance register in Figure 12 is the form such a record would take (Tabassi, 2023; ISO/IEC, 2023a; ISO/IEC, 2023c). The EU AI Act sets requirements for human oversight and transparency for high-risk systems, and whether an alert-prioritisation tool is high-risk depends on its use, for instance as a safety component in the operation of critical infrastructure; the Act's application dates are phased and have been the subject of proposals for amendment, so both the classification and the dates need to be checked with the current text and with counsel (European Parliament and Council, 2024b). The NIS 2 Directive extends incident-reporting duties to a wider set of operators, which may raise the cost of a missed incident and so the loss scale, the input with the largest Sobol index in this study (European Parliament and Council, 2022). Figure 12 sets out a four-phase roadmap (baseline and data, calibration and shadow running, prioritisation with explanation, governance and adaptation) with the governance bodies and enablers each phase needs; the timings in it are indicative.

Figure 12: Implementation roadmap for uncertainty-calibrated explainable alert prioritization: four phases from baseline and data through calibration and shadow running and prioritisation with explanation to governance and adaptation, with governance bodies and enablers



Implications for analytics practice

The study illustrates five practices. Evaluate on the period the model will face and not on the period it was trained on: drift raised the prevalence from 4.2% to 5.6% and degraded every calibrator, and the check can be run from the same label store (Sommer & Paxson, 2010; Arp et al., 2022). Report proper scores and reliability curves next to ECE, which depends on binning (Brier, 1950; Naeini et al., 2015; Vaicenavicius et al., 2019). Separate the three questions of ranking, probability and decision, because the tools that answer each are different and a gain on one does not carry to the others. Treat the human part of a pipeline as a component to be measured, with reliance on wrong and right advice reported next to accuracy (Bainbridge, 1983; Amershi et al., 2019). And propagate the decision into money under uncertainty with Sobol indices, so that the team knows which input to measure first: in this study the loss scale and the attack rate, not the explanation parameters (Sobol', 2001; Saltelli et al., 2010; Herman & Usher, 2017).

Limitations

Table 15 lists the limitations with their consequences and the way each could be addressed. The most important are three. The data, the drift, the costs and the analyst model are synthetic, so the orderings are conditional findings about a model world and absolute values, such as the ECE of 0.034 or the USD figures, carry no empirical weight. The analyst model has assumed parameters, and the effects of displays, flags and explanations are consequences of those assumptions. And labels are complete and correct, which real SOC labels are not; selection bias in labels is a known problem for models trained on reviewed alerts (Lakkaraju et al., 2017; Pendlebury et al., 2019). The study used three seeds and 20 test days, so intervals for rare families and for top-tier incidents are wide, and only differences much larger than the standard deviation (calibrated against raw values, display calibration, the doubling under the cost-optimal rule) should be taken as supported; the differences of about one thousand USD among the calibrated queues are not. Review is a single step without queue dynamics or fatigue (Harchol-Balter, 2013), explanations are TreeSHAP with an independence assumption that correlated features violate (Aas et al., 2021; Slack et al., 2020), and costs are expected values with no regulatory or reputational terms. Finally, the reference list was checked against publisher and registry records, but a small number of fields (page spans for some articles, print years inferred from online-first dates, and some proceedings details) could not be confirmed and should be checked before citation.

Further research

Five directions follow. First, replay on real alert histories: a SOC's labelled alerts with their review outcomes, a randomised audit sample of unreviewed alerts and measured loss tiers would replace the generators, and would test whether the orderings found here survive selective labels (Lakkaraju et al., 2017; Pendlebury et al., 2019). Second, studies with analysts, with the reliance measures of Table 13, to estimate the real effects of a calibrated probability, a flag and an explanation, including the cost of the extra review time (Bansal et al., 2021; Buçinca et al., 2021). Third, stronger detectors and richer uncertainty: sequence models on raw event streams, one-class and anomaly detectors for unseen attacks, and uncertainty methods designed to separate epistemic from aleatoric terms, with conformal methods designed for shift and delayed labels (Gawlikowski et al., 2023; Angelopoulos & Bates, 2023; Gibbs & Candès, 2021). Fourth, decision models beyond a single review step: learning to defer, queues with fatigue and shift change, and optimisation that is trained on the decision itself and not on prediction error (Madras et al., 2018; Mozannar & Sontag, 2020; Elmachtoub & Grigas, 2022). Fifth, explanations under adversaries and dependence: attacks on attributions, and attribution methods that respect correlated features (Slack et al., 2020; Aas et al., 2021).

Conclusion

This study examined, in a transparent simulation, whether calibrating alert scores, adding uncertainty signals and attaching explanations help prioritisation under limited analyst capacity. Four conclusions are supported within the limits of the testbed. A detector that ranks well can still be badly calibrated, and a one-parameter post-hoc calibrator repaired most of the error, with daily recalibration holding it under drift and parametric methods being safer with few labels. Calibration did not change a ranking by probability, but it raised the value of a queue by 17% when probability was multiplied by asset loss and about doubled the value of a cost-optimal review rule, which is the sense in which it matters. Uncertainty signals were less useful than they appear: conformal sets were valid and mostly

ambiguous, and ensemble spread was a risk proxy that did not signal drift and was unrelated to explanation stability. And under a stylised analyst model the calibrated probability was the display component that mattered, chiefly by avoiding the harm of a miscalibrated raw display, while flags and explanations exchanged recall for precision without a net gain. The Monte Carlo analysis shows that the size of these gains depends mostly on the loss scale and the attack rate, which an operator should measure before anything else. The analyst findings are hypotheses to be tested with people, the data are synthetic, and the numbers describe the testbed and not any real operation.

REFERENCES

[1]. Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence, 298, Article 103502. https://doi.org/10.1016/j.artint.2021.103502

[2]. Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76, 243–297. https://doi.org/10.1016/j.inffus.2021.05.008

[3]. Abdullah Mohammad, S. (2023). $CUGAO_2$ and $CUINS_2$ Quantum Dot Sensitization for Enhanced Photovoltaic Efficiency in Wearable Fiber-Shaped Solar Cells. American Journal of Advanced Technology and Engineering Solutions, 3(04), 209-261. https://doi.org/10.63125/wysdvy71

[4]. Abdullah Mohammad, S. (2024). Triboelectric And Piezoelectric Nanogenerators Using Binary and Ternary Oxide Nanostructures: Optimization of Energy Conversion Efficiency for Wearable Harvesting. Review of Applied Science and Technology, 3(03), 31–72. https://doi.org/10.63125/w7q9cs55

[5]. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[6]. Adams, N. M., & Hand, D. J. (1999). Comparing classifiers when the misallocation costs are uncertain. Pattern Recognition, 32(7), 1139–1147. https://doi.org/10.1016/S0031-3203(98)00154-X

[7]. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In Advances in Neural Information Processing Systems 31. Curran Associates. https://papers.nips.cc/paper_files/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html

[8]. Adnan, S., & Iftekhar, A. (2023). Predicting Cost Variance in Digital Transformation Projects Through Ensemble Learning and Project Performance Analytics. International Journal of Scientific Interdisciplinary Research, 13(10), 01–36. https://doi.org/10.63125/j18hxs51

[9]. Ahmed, H. C. H. R. (2024). Machine Learning Enhanced Demand Forecasting and Portfolio Analytics for Scalable Growth Management in Digital Logistics Operations. American Journal of Advanced Technology and Engineering Solutions, 4(02), 33-81. https://doi.org/10.63125/5xq3nd53

[10]. Ahmed, H. C. H. R., & Nuzhat Sadia, P. (2023). A Quantitative Assessment of CRM Analytics and KPI Dashboard Design on Enterprise Revenue Conversion in Emerging Market Logistics Platforms. American Journal of Data Science and Analytics, 4(09), 50-95. https://doi.org/10.63125/315bge15

[11]. Alahmadi, B. A., Axon, L., & Martinovic, I. (2022). 99% false positives: A qualitative study of SOC analysts' perspectives on security alarms. In Proceedings of the 31st USENIX Security Symposium (pp. 2783–2800). USENIX Association. https://www.usenix.org/conference/usenixsecurity22/presentation/alahmadi

[12]. Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods (1806.08049). arXiv. https://arxiv.org/abs/1806.08049

[13]. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19) (pp. 1–13). ACM. https://doi.org/10.1145/3290605.3300233

[14]. Aminul Islam, M. (2024). Interpreting Cyber-Threat Indicators in Enterprise Memory Forensics: A Mixed-Methods Study of Machine Learning Performance and Explainability. Review of Applied Science and Technology, 14(12), 01–13. https://doi.org/10.63125/2p8hte92

[15]. Aminul Islam, M., & Md, A. (2023). Adaptive Class-Imbalance Learning for High-Sensitivity Fraud Detection in Large-Scale Digital Transaction Networks. International Journal of Business and Economics Insights, 13(09), 01-33. https://doi.org/10.63125/bmxwv741

[16]. Ancker, J. S., Edwards, A., Nosal, S., Hauser, D., Mauer, E., & Kaushal, R. (2017). Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Medical Informatics and Decision Making, 17(1), Article 36. https://doi.org/10.1186/s12911-017-0430-8

[17]. Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning, 16(4), 494–591. https://doi.org/10.1561/2200000101

[18]. Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., & Schuster, T. (2024). Conformal risk control. In International Conference on Learning Representations (ICLR 2024). OpenReview. https://openreview.net/forum?id=33XGfHLtZg

[19]. Angelopoulos, A., Bates, S., Malik, J., & Jordan, M. I. (2021). Uncertainty sets for image classifiers using conformal prediction. In International Conference on Learning Representations (ICLR 2021). OpenReview. https://openreview.net/forum?id=eNdiU_DbM9

[20]. Arp, D., Quiring, E., Pendlebury, F., Warnecke, A., Pierazzi, F., Wressnegger, C., Cavallaro, L., & Rieck, K. (2022). Dos and don'ts of machine learning in computer security. In Proceedings of the 31st USENIX Security Symposium (pp. 3971–3988). USENIX Association. https://www.usenix.org/conference/usenixsecurity22/presentation/arp

[21]. Asheke, M. (2023). Machine Learning–Based Predictive Quality Modeling for Defect Detection and Process Optimization in Advanced Manufacturing Systems. Review of Applied Science and Technology, 13(12), 01–49. https://doi.org/10.63125/thy73b16

[22]. Asheke, M. (2024). Digital Twin and Industrial Internet of Things–Enabled Predictive Maintenance for Reliability Enhancement in Smart Manufacturing Systems. International Journal of Scientific Interdisciplinary Research, 5(2), 792–838. https://doi.org/10.63125/jgrdqd62

[23]. Ashik, M., & Shabbir, A. (2022). Digital Twin-Enabled Predictive Maintenance for Rotating and Static Equipment: Bridging FEA-Based Failure Analysis with Real-Time Asset Health Monitoring in Energy and Power Facilities. International Journal of Scientific Interdisciplinary Research, 12(05), 01–37. https://doi.org/10.63125/0qaysz49

[24]. Ashrafuzzaman, M. (2024). The impact of cloud-based management information systems on hrm efficiency: an analysis of small and medium-sized enterprises (SMES). Academic Journal on Artificial Intelligence, Machine Learning, Data Science and Management Information Systems, 1(01), 40-56. https://doi.org/10.69593/ajaimldsmis.v1i01.124

[25]. Axelsson, S. (2000). The base-rate fallacy and the difficulty of intrusion detection. ACM Transactions on Information and System Security, 3(3), 186–205. https://doi.org/10.1145/357830.357849

[26]. Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

[27]. Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019). Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

[28]. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21) (pp. 1–16). ACM. https://doi.org/10.1145/3411764.3445717

[29]. Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. The Annals of Statistics, 51(2). https://doi.org/10.1214/23-AOS2276

[30]. Barbero, F., Pendlebury, F., Pierazzi, F., & Cavallaro, L. (2022). Transcending Transcend: Revisiting malware classification in the presence of concept drift. In 2022 IEEE Symposium on Security and Privacy (SP) (pp. 805–823). IEEE. https://doi.org/10.1109/SP46214.2022.9833659

[31]. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[32]. Bastani, O., Kim, C., & Bastani, H. (2017). Interpretability via model extraction (1706.09773). arXiv. https://arxiv.org/abs/1706.09773

[33]. Bates, S., Angelopoulos, A., Lei, L., Malik, J., & Jordan, M. (2021). Distribution-free, risk-controlling prediction sets. Journal of the ACM, 68(6), Article 43. https://doi.org/10.1145/3478535

[34]. Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Srikumar, M., Weller, A., & Xiang, A. (2021). Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21) (pp. 401–413). ACM. https://doi.org/10.1145/3461702.3462571

[35]. Bhatt, U., Weller, A., & Moura, J. M. F. (2020). Evaluating and aggregating feature-based model explanations. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) (pp. 3016–3022). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2020/417

[36]. Breznitz, S. (1984). Cry wolf: The psychology of false alarms. Lawrence Erlbaum Associates. https://ci.nii.ac.jp/ncid/BA01002810

[37]. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

[38]. Bröcker, J., & Smith, L. A. (2007). Increasing the reliability of reliability diagrams. Weather and Forecasting, 22(3), 651–661. https://doi.org/10.1175/WAF993.1

[39]. Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1–21. https://doi.org/10.1145/3449287

[40]. Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). Explainable artificial intelligence in cybersecurity: A survey. IEEE Access, 10, 93575–93600. https://doi.org/10.1109/ACCESS.2022.3204171

[41]. Chow, C. K. (1970). On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory, 16(1), 41–46. https://doi.org/10.1109/TIT.1970.1054406

[42]. Cichonski, P., Millar, T., Grance, T., & Scarfone, K. (2012). Computer security incident handling guide (NIST Special Publication 800-61 Rev. 2). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-61r2

[43]. Cobham, A. (1954). Priority assignment in waiting line problems. Journal of the Operations Research Society of America, 2(1), 70–76. https://doi.org/10.1287/opre.2.1.70

[44]. Covert, I., Lundberg, S., & Lee, S.-I. (2021). Explaining by removing: A unified framework for model explanation. Journal of Machine Learning Research, 22(209), 1–90. https://jmlr.org/papers/v22/20-1316.html

[45]. Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. In AIAA 1st Intelligent Systems Technical Conference. American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2004-6313

[46]. Cuppens, F., & Miège, A. (2002). Alert correlation in a cooperative intrusion detection framework. In Proceedings 2002 IEEE Symposium on Security and Privacy (pp. 202–215). IEEE. https://doi.org/10.1109/SECPRI.2002.1004372

[47]. Cvach, M. (2012). Monitor alarm fatigue: An integrative review. Biomedical Instrumentation & Technology, 46(4), 268–277. https://doi.org/10.2345/0899-8205-46.4.268

[48]. Dawid, A. P. (1982). The well-calibrated Bayesian. Journal of the American Statistical Association, 77(379), 605–610. https://doi.org/10.1080/01621459.1982.10477856

[49]. DeGroot, M. H., & Fienberg, S. E. (1983). The comparison and evaluation of forecasters. The Statistician, 32(1/2), 12–22. https://doi.org/10.2307/2987588

[50]. Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., & Udluft, S. (2018). Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In Proceedings of the 35th International Conference on Machine Learning (ICML), PMLR 80 (pp. 1184–1193). PMLR. https://proceedings.mlr.press/v80/depeweg18a.html

[51]. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126. https://doi.org/10.1037/xge0000033

[52]. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning (arXiv:1702.08608). arXiv. https://arxiv.org/abs/1702.08608

[53]. Drummond, C., & Holte, R. C. (2006). Cost curves: An improved method for visualizing classifier performance. Machine Learning, 65(1), 95–130. https://doi.org/10.1007/s10994-006-8199-5

[54]. Elkan, C. (2001). The foundations of cost-sensitive learning. In Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI '01) (pp. 973–978). Morgan Kaufmann. https://mlanthology.org/ijcai/2001/elkan2001ijcai-foundations

[55]. Elmachtoub, A. N., & Grigas, P. (2022). Smart "predict, then optimize". Management Science, 68(1), 9–26. https://doi.org/10.1287/mnsc.2020.3922

[56]. El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. Journal of Machine Learning Research, 11, 1605–1641. https://jmlr.org/papers/v11/el-yaniv10a.html

[57]. Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37(1), 32–64. https://doi.org/10.1518/001872095779049543

[58]. Engelen, G., Rimmer, V., & Joosen, W. (2021). Troubleshooting an intrusion detection dataset: The CICIDS2017 case study. In 2021 IEEE Security and Privacy Workshops (SPW) (pp. 7–12). IEEE. https://doi.org/10.1109/SPW53761.2021.00009

[59]. European Parliament and Council. (2022). Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union (NIS 2 Directive). Official Journal of the European Union, L 333, 80-152. https://eur-lex.europa.eu/eli/dir/2022/2555/oj

[60]. European Parliament and Council. (2024b). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[61]. Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research, 20(177), 1–81. https://jmlr.org/papers/v20/18-760.html

[62]. Fort, S., Hu, H., & Lakshminarayanan, B. (2019). Deep ensembles: A loss landscape perspective (1912.02757). arXiv. https://arxiv.org/abs/1912.02757

[63]. Foygel Barber, R., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2021). The limits of distribution-free conditional predictive inference. Information and Inference: A Journal of the IMA, 10(2), 455–482. https://doi.org/10.1093/imaiai/iaaa017

[64]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[65]. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (PMLR 48) (pp. 1050–1059). https://proceedings.mlr.press/v48/gal16.html

[66]. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), Article 44. https://doi.org/10.1145/2523813

[67]. Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., & Zhu, X. X. (2023). A survey of uncertainty in deep neural networks. Artificial Intelligence Review, 56(Suppl. 1), 1513–1589. https://doi.org/10.1007/s10462-023-10562-9

[68]. Gazi Yeasin, A. (2023). Impact of Accessibility-First Design Systems on WCAG 2.1 Compliance and Inclusive User Experience in Enterprise Software Platforms: A Mixed-Methods Evaluation. Review of Applied Science and Technology, 13(07), 01–49. https://doi.org/10.63125/nxh4gp19

[69]. Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In Advances in Neural Information Processing Systems 30. Curran Associates. https://papers.nips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html

[70]. Geifman, Y., & El-Yaniv, R. (2019). SelectiveNet: A deep neural network with an integrated reject option. In Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97 (pp. 2151–2159). PMLR. https://proceedings.mlr.press/v97/geifman19a.html

[71]. Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. In Proceedings of the AAAI Conference on Artificial Intelligence, 33(1) (pp. 3681–3688). AAAI Press. https://doi.org/10.1609/aaai.v33i01.33013681

[72]. Gibbs, I., & Candès, E. (2021). Adaptive conformal inference under distribution shift. In Advances in Neural Information Processing Systems 34. Curran Associates. https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html

[73]. Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359–378. https://doi.org/10.1198/016214506000001437

[74]. Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association, 19(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

[75]. Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–24. https://doi.org/10.1145/3359152

[76]. Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. MIS Quarterly, 37(2), 337–355. https://doi.org/10.25300/MISQ/2013/37.2.01

[77]. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), Article 93. https://doi.org/10.1145/3236009

[78]. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (PMLR 70, pp. 1321–1330). https://proceedings.mlr.press/v70/guo17a.html

[79]. Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. Machine Learning, 77(1), 103–123. https://doi.org/10.1007/s10994-009-5119-5

[80]. Harchol-Balter, M. (2013). Performance modeling and design of computer systems: Queueing theory in action. Cambridge University Press. https://doi.org/10.1017/CBO9781139226424

[81]. Hassan, W. U., Bates, A., & Marino, D. (2020). Tactical provenance analysis for endpoint detection and response systems. In 2020 IEEE Symposium on Security and Privacy (SP) (pp. 1172–1189). IEEE. https://doi.org/10.1109/SP40000.2020.00096

[82]. Hassan, W. U., Guo, S., Li, D., Chen, Z., Jee, K., Li, Z., & Bates, A. (2019). NoDoze: Combatting threat alert fatigue with automated provenance triage. In Proceedings of the 2019 Network and Distributed System Security Symposium (NDSS). Internet Society. https://doi.org/10.14722/ndss.2019.23349

[83]. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

[84]. Hébert-Johnson, Ú., Kim, M. P., Reingold, O., & Rothblum, G. N. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. In Proceedings of the 35th International Conference on Machine Learning (ICML), PMLR 80 (pp. 1939–1948). PMLR. https://proceedings.mlr.press/v80/hebert-johnson18a.html

[85]. Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In International Conference on Learning Representations (ICLR 2017). OpenReview. https://openreview.net/forum?id=Hkg4TI9xl

[86]. Herman, J., & Usher, W. (2017). SALib: An open-source Python library for sensitivity analysis. Journal of Open Source Software, 2(9), Article 97. https://doi.org/10.21105/joss.00097

[87]. Hernández-Orallo, J., Flach, P., & Ferri, C. (2012). A unified view of performance metrics: Translating threshold choice into expected classification loss. Journal of Machine Learning Research, 13, 2813–2869. https://jmlr.org/papers/v13/hernandez-orallo12a.html

[88]. Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105. https://doi.org/10.2307/25148625

[89]. Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 57(3), 407–434. https://doi.org/10.1177/0018720814547570

[90]. Hooker, S., Erhan, D., Kindermans, P.-J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. In Advances in Neural Information Processing Systems 32. Curran Associates. https://papers.nips.cc/paper_files/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html

[91]. Horvitz, E. (1999). Principles of mixed-initiative user interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99) (pp. 159–166). ACM. https://doi.org/10.1145/302979.303030

[92]. Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning, 110(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3

[93]. ISO/IEC. (2023a). ISO/IEC 42001:2023 Information technology: Artificial intelligence: Management system. ISO. https://www.iso.org/standard/81230.html

[94]. ISO/IEC. (2023b). ISO/IEC 27035-1:2023 Information technology: Information security incident management: Part 1: Principles and process. ISO. https://www.iso.org/standard/78973.html

[95]. ISO/IEC. (2023c). ISO/IEC 23894:2023 Information technology: Artificial intelligence: Guidance on risk management. ISO. https://www.iso.org/standard/77304.html

[96]. Jahanara, A., & Mst Kaniz, F. (2023). Demand Variability and Capacity Planning in Home and Community-Based Services: A Decision-Analytic Framework for Health Services Managers. American Journal of Data Science and Analytics, 4(09), 01-49. https://doi.org/10.63125/3y62vr16

[97]. Jannatul, F. (2024). Alzheimer's Disease and Mobility Multimodal Prediction of Gait Decline, Fall Risk, and Loss of Functional Independence in Mild Cognitive Impairment and Early Alzheimer's Disease. Review of Applied Science and Technology, 3(01), 420–483. https://doi.org/10.63125/q2d90m63

[98]. Janzing, D., Minorics, L., & Blöbaum, P. (2020). Feature relevance quantification in explainable AI: A causal problem. In Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR 108 (pp. 2907–2916). PMLR. https://proceedings.mlr.press/v108/janzing20a.html

[99]. Jordaney, R., Sharad, K., Dash, S. K., Wang, Z., Papini, D., Nouretdinov, I., & Cavallaro, L. (2017). Transcend: Detecting concept drift in malware classification models. In Proceedings of the 26th USENIX Security Symposium (pp. 625–642). USENIX Association. https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/jordaney

[100]. Julisch, K. (2003). Clustering intrusion detection alarms to support root cause analysis. ACM Transactions on Information and System Security, 6(4), 443–471. https://doi.org/10.1145/950191.950192

[101]. Julisch, K., & Dacier, M. (2002). Mining intrusion detection alarms for actionable knowledge. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02) (pp. 366–375). ACM. https://doi.org/10.1145/775047.775101

[102]. Kamar, E., Hacker, S., & Horvitz, E. (2012). Combining human and machine intelligence in large-scale crowdsourcing. In Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012). IFAAMAS. https://www.microsoft.com/en-us/research/publication/combining-human-and-machine-intelligence-in-large-scale-crowdsourcing/

[103]. Kazi Rakib Hasan, S., & Adnan, S. (2022). Quantitative Assessment of Machine Learning-Based Schedule Risk Prediction for Reducing Delays in Complex Technology Projects. Review of Applied Science and Technology, 1(04), 464–501. https://doi.org/10.63125/dn8nkm85

[104]. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154. https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree

[105]. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems 30 (NeurIPS 2017). Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html

[106]. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. The Quarterly Journal of Economics, 133(1), 237–293. https://doi.org/10.1093/qje/qjx032

[107]. Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19) (pp. 1–14). ACM. https://doi.org/10.1145/3290605.3300641

[108]. Kokulu, F. B., Soneji, A., Bao, T., Shoshitaishvili, Y., Zhao, Z., Doupé, A., & Ahn, G.-J. (2019). Matched and mismatched SOCs: A qualitative study on security operations center issues. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19) (pp. 1955–1970). ACM. https://doi.org/10.1145/3319535.3354239

[109]. Krishna, S., Han, T., Gu, A., Wu, S., Jabbari, S., & Lakkaraju, H. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective (2202.01602). arXiv. https://arxiv.org/abs/2202.01602

[110]. Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., & Flach, P. (2019). Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. In Advances in Neural Information Processing Systems 32. Curran Associates. https://papers.nips.cc/paper_files/paper/2019/hash/8ca01ea920679a0fe3728441494041b9-Abstract.html

[111]. Kull, M., Silva Filho, T., & Flach, P. (2017). Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR 54 (pp. 623–631). PMLR. https://proceedings.mlr.press/v54/kull17a.html

[112]. Kumar, A., Liang, P., & Ma, T. (2019). Verified uncertainty calibration. In Advances in Neural Information Processing Systems 32. Curran Associates. https://papers.nips.cc/paper_files/paper/2019/hash/f8c0c968632845cd133308b1a494967f-Abstract.html

[113]. Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. In Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR 119 (pp. 5491–5500). PMLR. https://proceedings.mlr.press/v119/kumar20e.html

[114]. Kynatun, M., & Saba, A. (2024). Machine Learning Enhanced Intrusion Detection and Anomaly Classification for MIS-Driven Enterprise Network Security. International Journal of Scientific Interdisciplinary Research, 14(09), 01–41. https://doi.org/10.63125/xk6kpc72

[115]. Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19) (pp. 29–38). ACM. https://doi.org/10.1145/3287560.3287590

[116]. Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17) (pp. 275–284). ACM. https://doi.org/10.1145/3097983.3098066

[117]. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017). https://arxiv.org/abs/1612.01474

[118]. Law, A. M. (2015). Simulation modeling and analysis (5th ed.). McGraw-Hill Education. https://www.mheducation.com/

[119]. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

[120]. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. Journal of the American Statistical Association, 113(523), 1094–1111. https://doi.org/10.1080/01621459.2017.1307116

[121]. Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31–57. https://doi.org/10.1145/3236386.3241340

[122]. Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

[123]. Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2019). Learning under concept drift: A review. IEEE Transactions on Knowledge and Data Engineering, 31(12), 2346–2363. https://doi.org/10.1109/TKDE.2018.2876857

[124]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

[125]. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[126]. Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., & Wilson, A. G. (2019). A simple baseline for Bayesian uncertainty in deep learning. In Advances in Neural Information Processing Systems 32. Curran Associates. https://papers.neurips.cc/paper_files/paper/2019/file/118921efba23fc329e6560b27861f0c2-Paper.pdf

[127]. Madras, D., Pitassi, T., & Zemel, R. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. In Advances in Neural Information Processing Systems 31. Curran Associates. https://papers.nips.cc/paper_files/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html

[128]. Mahbooba, B., Timilsina, M., Sahal, R., & Serrano, M. (2021). Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. Complexity, 2021, Article 6634811. https://doi.org/10.1155/2021/6634811

[129]. Mansura Akter, E., & Sumaiya, M. (2022). Antimicrobial Resistance Profiles of Acinetobacter Baumannii and Vibrio Cholerae in Hospital Wastewater and Adjacent Household Water Sources. American Journal of Data Science and Analytics, 3(09), 01-55. https://doi.org/10.63125/gbervt14

[130]. Mashfiqur, R., & Md, A. (2022). Monte Carlo Simulation and Stochastic Cash-Flow Modeling for Reducing Liquidity Risk in Project-Based Enterprises. International Journal of Scientific Interdisciplinary Research, 1(01), 382–431. https://doi.org/10.63125/zf3fwv21

[131]. Mashiat, N. (2024). Reducing Telecommunications Service Interruptions Through Predictive Root-Cause Analytics of Multivariate System Telemetry. International Journal of Scientific Interdisciplinary Research, 14(11), 01-37. https://doi.org/10.63125/nqjh0c29

[132]. Maslach, C., Schaufeli, W. B., & Leiter, M. P. (2001). Job burnout. Annual Review of Psychology, 52(1), 397–422. https://doi.org/10.1146/annurev.psych.52.1.397

[133]. McHugh, J. (2000). Testing intrusion detection systems: A critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. ACM Transactions on Information and System Security, 3(4), 262–294. https://doi.org/10.1145/382912.382923

[134]. Md Ashfaq, S., & Abdullah Mohammad, S. (2022). CNT/PANI Composite Electrodes for Flexible Fiber-Shaped Dye-Sensitized Solar Cells: Synthesis, Characterization, and Photovoltaic Performance. American Journal of Interdisciplinary Studies, 3(02), 221-274. https://doi.org/10.63125/m334mf41

[135]. Md Asif Ali Sheak, A., & Tahmedur, R. (2023). A Data-Driven Framework for Engineering Project Risk Assessment Under Schedule and Cost Uncertainty. International Journal of Scientific Interdisciplinary Research, 4(2), 204–251. https://doi.org/10.63125/kq4zzh16

[136]. Md Imran, H., & Jannatul, F. (2021). Translational Stroke and Neurorehabilitation Research Integrating Neuroimaging, Circulating Biomarkers, and Functional Assessments to Predict Recovery After Stroke. American

Journal of Data Science and Analytics, 2(12), 156-198. https://doi.org/10.63125/c28rzv23

[137]. Md Imran, H., & Jannatul, F. (2022). Molecular and Precision Rehabilitation Molecular Mechanisms Linking Mitochondrial Dysfunction, Inflammation, and Muscle Degeneration with Functional Decline and Rehabilitation Response in Neuromuscular Disease. American Journal of Interdisciplinary Studies, 12(08), 01-44. https://doi.org/10.63125/0ym5bh79

[138]. Md Kaium, H. (2023). A GIS-Based Multi-Criteria Decision-Support and Land-Suitability Optimization Framework for Climate-Resilient Urban and Regional Planning. American Journal of Data Science and Analytics, 4(06), 43–103. https://doi.org/10.63125/wvq0sx87

[139]. Md Mahamud Pervaz, P. (2023). Integration of MES, SAP, And PI Historian Platforms for Real-Time Industrial Asset Performance Management. Review of Applied Science and Technology, 2(04), 393–445. https://doi.org/10.63125/9ebtas97

[140]. Md, A., & Gazi Yeasin, A. (2022). Machine Learning Enhanced Spatial-Data Workflow Redesign: A Quantitative Evaluation of Task Success and Operational Cost Reduction in Enterprise LiDAR Platforms. American Journal of Data Science and Analytics, 3(06), 80-136. https://doi.org/10.63125/0den7c02

[141]. Md. Morshedul, I., & Aminul Islam, M. (2022). A Quantitative Evaluation of Explainable Ensemble Learning for Probabilistic Risk Estimation and Decision Reliability in Enterprise Analytics. American Journal of Data Science and Analytics, 3(02), 01-38. https://doi.org/10.63125/k22shk74

[142]. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[143]. Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., & Lucic, M. (2021). Revisiting the calibration of modern neural networks. In Advances in Neural Information Processing Systems 34. Curran Associates. https://proceedings.neurips.cc/paper/2021/hash/8420d359404024567b5aefda1231af24-Abstract.html

[144]. Mohammad Badrul, A., & Md. Mominul, H. (2023). Machine Learning–Based Assessment of Environmental and Public Health Risks in Sewerage Sludge Management Systems. American Journal of Advanced Technology and Engineering Solutions, 3(02), 33-77. https://doi.org/10.63125/y3yfxa92

[145]. Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. Pattern Recognition, 45(1), 521–530. https://doi.org/10.1016/j.patcog.2011.06.019

[146]. Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20) (pp. 607–617). ACM. https://doi.org/10.1145/3351095.3372850

[147]. Mozannar, H., & Sontag, D. (2020). Consistent estimators for learning to defer to an expert. In Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR 119 (pp. 7076–7087). PMLR. https://proceedings.mlr.press/v119/mozannar20b.html

[148]. Mst Kaniz, F., & Jahanara, A. (2022). Lean Six Sigma Implementation in Home Health Administrative Operations: A DMAIC Case Study of Managerial Workflow Redesign and Error Reduction. International Journal of Scientific Interdisciplinary Research, 12(04), 01–60. https://doi.org/10.63125/xchca204

[149]. Murphy, A. H. (1973). A new vector partition of the probability score. Journal of Applied Meteorology, 12(4), 595–600. https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2

[150]. Nadeem, A., Vos, D., Cao, C., Pajola, L., Dieck, S., Baumgartner, R., & Verwer, S. (2023). SoK: Explainable machine learning for computer security applications. In 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P) (pp. 221–240). IEEE. https://doi.org/10.1109/EuroSP57164.2023.00022

[151]. Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (pp. 2901–2907). AAAI Press. https://doi.org/10.1609/aaai.v29i1.9602

[152]. National Institute of Standards and Technology. (2024). The NIST Cybersecurity Framework (CSF) 2.0 (NIST CSWP 29). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.CSWP.29

[153]. Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., & Seale, M. (2022). Explainable intrusion detection systems (X-IDS): A survey of current methods, challenges, and opportunities (2207.06236). arXiv. https://arxiv.org/abs/2207.06236

[154]. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning (pp. 625–632). ACM. https://doi.org/10.1145/1102351.1102430

[155]. Nishat, J. (2023). Communicable Disease Prevention in a Post-Pandemic Era: A Global Review of Public Health Education, Surveillance, and Community-Level Interventions. International Journal of Scientific Interdisciplinary Research, 4(4), 630–678. https://doi.org/10.63125/mnqefj36

[156]. Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., & Tran, D. (2019). Measuring calibration in deep learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (pp. 38–41) IEEE/CVF. https://openaccess.thecvf.com/content_CVPRW_2019/html/Uncertainty_and_Robustness_in_Deep_Visual_Learning/Nixon_Measuring_Calibration_in_Deep_Learning_CVPRW_2019_paper.html

[157]. Nuzhat Sadia, P. (2023). A Multi-Criteria Decision-Support Framework For ERP-Based Business Process Optimization in Mid-Sized Manufacturing Enterprises. American Journal of Scholarly Research and Innovation, 2(02), 519–573. https://doi.org/10.63125/j58k5d55

[158]. Okati, N., De, A., & Gomez-Rodriguez, M. (2021). Differentiable learning under triage. In Advances in Neural Information Processing Systems 34. Curran Associates. https://arxiv.org/abs/2103.08902

[159]. Oprea, A., Li, Z., Yen, T.-F., Chin, S. H., & Alrwais, S. (2015). Detection of early-stage enterprise infection by mining large-scale log data. In 2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (pp. 45–56). IEEE. https://doi.org/10.1109/DSN.2015.14

[160]. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In Advances in Neural Information Processing Systems 32. Curran Associates. https://papers.nips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html

[161]. Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. (2002). Inductive confidence machines for regression. In Machine Learning: ECML 2002, Lecture Notes in Computer Science 2430 (pp. 345–356). Springer. https://doi.org/10.1007/3-540-36755-1_29

[162]. Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410. https://doi.org/10.1177/0018720810376055

[163]. Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253. https://doi.org/10.1518/001872097778543886

[164]. Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 30(3), 286–297. https://doi.org/10.1109/3468.844354

[165]. Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

[166]. Pendlebury, F., Pierazzi, F., Jordaney, R., Kinder, J., & Cavallaro, L. (2019). TESSERACT: Eliminating experimental bias in malware classification across space and time. In Proceedings of the 28th USENIX Security Symposium (pp. 729–746). USENIX Association. https://www.usenix.org/conference/usenixsecurity19/presentation/pendlebury

[167]. Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. (2021). Four principles of explainable artificial intelligence (NISTIR 8312). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8312

[168]. Pietraszek, T. (2004). Using adaptive alert classification to reduce false positives in intrusion detection. In Recent Advances in Intrusion Detection (RAID 2004), Lecture Notes in Computer Science 3224 (pp. 102–124). Springer. https://doi.org/10.1007/978-3-540-30143-1_6

[169]. Pinedo, M. L. (2016). Scheduling: Theory, algorithms, and systems (5th ed.). Springer. https://doi.org/10.1007/978-3-319-26580-3

[170]. Platt, J. C. (2000). Probabilities for SV machines. In A. J. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), Advances in large margin classifiers (pp. 61–74). MIT Press. https://doi.org/10.7551/mitpress/1113.003.0008

[171]. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In Advances in Neural Information Processing Systems 30 (NIPS 2017). Curran Associates. https://papers.nips.cc/paper/7151-on-fairness-and-calibration

[172]. Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J., & Wallach, H. (2021). Manipulating and measuring model interpretability. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21) (pp. 1–52). ACM. https://doi.org/10.1145/3411764.3445315

[173]. Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. Machine Learning, 42(3), 203–231. https://doi.org/10.1023/A:1007601015854

[174]. Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds.). (2009). Dataset shift in machine learning. MIT Press. https://doi.org/10.7551/mitpress/9780262170055.001.0001

[175]. Raghu, M., Blumer, K., Corrado, G., Kleinberg, J., Obermeyer, Z., & Mullainathan, S. (2019). The algorithmic automation problem: Prediction, triage, and human effort (1903.12220). arXiv. https://arxiv.org/abs/1903.12220

[176]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

[177]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 1527–1535. https://doi.org/10.1609/aaai.v32i1.11491

[178]. Romano, Y., Sesia, M., & Candès, E. (2020). Classification with valid and adaptive coverage. In Advances in Neural Information Processing Systems 33. Curran Associates. https://proceedings.neurips.cc/paper/2020/hash/244edd7e85dc81602b7615cd705545f5-Abstract.html

[179]. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[180]. Sadia, Z., & Maishar Mostari, Z. (2022). The Impact of Interactive Dashboard Visualization on Data-Driven Decision Making in Financial Services: A Qualitative Evaluation. International Journal of Scientific Interdisciplinary Research, 12(03), 01–35. https://doi.org/10.63125/3jstsv75

[181]. Sadinle, M., Lei, J., & Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. Journal of the American Statistical Association, 114(525), 223–234. https://doi.org/10.1080/01621459.2017.1395341

[182]. Saerens, M., Latinne, P., & Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. Neural Computation, 14(1), 21–41. https://doi.org/10.1162/089976602753284446

[183]. Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., & Tarantola, S. (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Computer Physics Communications, 181(2), 259–270. https://doi.org/10.1016/j.cpc.2009.09.018

[184]. Sargent, R. G. (2013). Verification and validation of simulation models. Journal of Simulation, 7(1), 12–24. https://doi.org/10.1057/jos.2012.20

[185]. Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. In Advances in Neural Information Processing Systems 31. Curran Associates. https://proceedings.neurips.cc/paper/2018/hash/a981f2b708044d6fb4a71a1463242520-Abstract.html

[186]. Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. Journal of Machine Learning Research, 9, 371–421. https://jmlr.org/papers/v9/shafer08a.html

[187]. Shapley, L. S. (1953). A value for n-person games. In Contributions to the Theory of Games (AM-28), Volume II (pp. 307–318). Princeton University Press. https://doi.org/10.1515/9781400881970-018

[188]. Skitka, L. J., Mosier, K. L., & Burdick, M. D. (1999). Does automation bias decision-making? International Journal of Human-Computer Studies, 51(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

[189]. Slack, D., Hilgard, A., Singh, S., & Lakkaraju, H. (2021). Reliable post hoc explanations: Modeling uncertainty in explainability. In Advances in Neural Information Processing Systems 34. Curran Associates. https://neurips.cc/virtual/2021/poster/27871

[190]. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20) (pp. 180–186). ACM. https://doi.org/10.1145/3375627.3375830

[191]. Smith, W. E. (1956). Various optimizers for single-stage production. Naval Research Logistics Quarterly, 3(1-2), 59–66. https://doi.org/10.1002/nav.3800030106

[192]. Sobol', I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and Computers in Simulation, 55(1-3), 271–280. https://doi.org/10.1016/S0378-4754(00)00270-6

[193]. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In 2010 IEEE Symposium on Security and Privacy (pp. 305–316). IEEE. https://doi.org/10.1109/SP.2010.25

[194]. Spathoulas, G. P., & Katsikas, S. K. (2010). Reducing false positives in intrusion detection systems. Computers & Security, 29(1), 35–44. https://doi.org/10.1016/j.cose.2009.07.008

[195]. Stabak, D., & Md Shahid, N. (2022). Empirical Assessment of Uncertainty-Aware Real-Time Decision-Making for Reliable and Resilient Cyber-Physical Systems. International Journal of Scientific Interdisciplinary Research, 1(01), 432–477. https://doi.org/10.63125/q3r4kk16

[196]. Stabak, D., & Md Siam, T. (2023). Real-Time Edge Computing for Cyber-Physical Systems: A Meta-Analysis of Energy Efficiency and Fault Tolerance. American Journal of Scholarly Research and Innovation, 13(06), 01–44. https://doi.org/10.63125/g9c9cp41

[197]. Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems, 41(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x

[198]. Sumaiya, M. (2024). Hospital Wastewater as A Reservoir of Antimicrobial Resistance Genes: Implications for Public Health and Genomic Surveillance in The United States. Review of Applied Science and Technology, 3(04), 448–508. https://doi.org/10.63125/ahvas976

[199]. Sundaramurthy, S. C., Bardas, A. G., Case, J., Ou, X., Wesch, M., McHugh, J., & Rajagopalan, S. R. (2015). A human capital model for mitigating security analyst burnout. In Eleventh Symposium On Usable Privacy and Security (SOUPS 2015) (pp. 347–359). USENIX Association. https://www.usenix.org/conference/soups2015/proceedings/presentation/sundaramurthy

[200]. Syed Feroz, A. (2023). Quality-Management Systems for Resilient Energy Infrastructure Fabrication. International Journal of Scientific Interdisciplinary Research, 13(10), 37–91. https://doi.org/10.63125/mbhhrm88

[201]. Syed Feroz, A. (2024). AI-Assisted NDT Analytics for Early Crack Detection in Energy-Infrastructure Pressure Vessels Using Advanced Manufacturing Data. Review of Applied Science and Technology, 14(07), 01-52. https://doi.org/10.63125/8bdqs538

[202]. Syed Feroz, A., & Md Mahamud Pervaz, P. (2022). Root-Cause Analysis and FMEA-Driven Reliability Improvement in Gas Production Facilities: A Multi-Year Empirical Study. International Journal of Scientific Interdisciplinary Research, 12(10), 01–53. https://doi.org/10.63125/70msp878

[203]. Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

[204]. Tahmedur, R., & Rifat, C. (2023). Critical Path Analysis and Resource-Leveling Optimization for Multi-Project Engineering Portfolios. American Journal of Interdisciplinary Studies, 13(12), 01-47. https://doi.org/10.63125/8eqnea46

[205]. Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications (pp. 1–6). IEEE. https://doi.org/10.1109/CISDA.2009.5356528

[206]. Tibshirani, R. J., Foygel Barber, R., Candès, E., & Ramdas, A. (2019). Conformal prediction under covariate shift. In Advances in Neural Information Processing Systems 32. Curran Associates. https://papers.nips.cc/paper/8522-conformal-prediction-under-covariate-shift

[207]. Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., & Schön, T. B. (2019). Evaluating model calibration in classification. In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR 89 (pp. 3459–3467). PMLR. https://proceedings.mlr.press/v89/vaicenavicius19a.html

[208]. Valdes, A., & Skinner, K. (2001). Probabilistic alert correlation. In Recent Advances in Intrusion Detection (RAID 2001), Lecture Notes in Computer Science 2212 (pp. 54–68). Springer. https://doi.org/10.1007/3-540-45474-8_4

[209]. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, 230. https://doi.org/10.1186/s12916-019-1466-7

[210]. Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1–38. https://doi.org/10.1145/3579605

[211]. Veeramachaneni, K., Arnaldo, I., Korrapati, V., Bassias, C., & Li, K. (2016). AI²: Training a big data machine to defend. In 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS) (pp. 49–54). IEEE. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2016.79

[212]. Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making, 26(6), 565–574. https://doi.org/10.1177/0272989X06295361

[213]. Vielberth, M., Böhm, F., Fichtinger, I., & Pernul, G. (2020). Security operations center: A systematic study and open challenges. IEEE Access, 8, 227756–227779. https://doi.org/10.1109/ACCESS.2020.3045514

[214]. Vovk, V. (2012). Conditional validity of inductive conformal predictors. In Proceedings of the Asian Conference on Machine Learning (ACML), PMLR 25 (pp. 475–490). PMLR. https://proceedings.mlr.press/v25/vovk12.html

[215]. Vovk, V., Gammerman, A., & Shafer, G. (2005). Algorithmic learning in a random world. Springer. https://doi.org/10.1007/b106715

[216]. Wang, M., Zheng, K., Yang, Y., & Wang, X. (2020). An explainable machine learning framework for intrusion detection systems. IEEE Access, 8, 73127–73141. https://doi.org/10.1109/ACCESS.2020.2988359

[217]. Wang, X., & Yin, M. (2021). Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21) (pp. 318–328). ACM. https://doi.org/10.1145/3397481.3450650

[218]. Widmann, D., Lindsten, F., & Zachariah, D. (2019). Calibration tests in multi-class classification: A unifying framework. In Advances in Neural Information Processing Systems 32. Curran Associates. https://proceedings.neurips.cc/paper/2019/hash/1c336b8080f82bcc2cd2499b4c57261d-Abstract.html

[219]. Wilder, B., Horvitz, E., & Kamar, E. (2020). Learning to complement humans. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) (pp. 1526–1533). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2020/212

[220]. Wilson, A. G., & Izmailov, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. In Advances in Neural Information Processing Systems 33. Curran Associates. https://papers.neurips.cc/paper_files/paper/2020/file/322f62469c5e3c7dc3e58f5a4d1ea399-Paper.pdf

[221]. Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., & Ravikumar, P. K. (2019). On the (in)fidelity and sensitivity of explanations. In Advances in Neural Information Processing Systems 32. Curran Associates. https://papers.nips.cc/paper_files/paper/2019/hash/a7471fdc77b3435276507cc8f2dc2569-Abstract.html

[222]. Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19) (pp. 1–12). ACM. https://doi.org/10.1145/3290605.3300509

[223]. Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02) (pp. 694–699). ACM. https://doi.org/10.1145/775047.775151

[224]. Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20) (pp. 295–305). ACM. https://doi.org/10.1145/3351095.3372852

[225]. Zhao, X., Huang, W., Huang, X., Robu, V., & Flynn, D. (2021). BayLIME: Bayesian local interpretable model-agnostic explanations. In Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR 161 (pp. 887–896). PMLR. https://proceedings.mlr.press/v161/zhao21a.html