International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(12s) 2026 pp. 89-96
https://doi.org/10.70917/ijcisim-2026-3880
Article

Confidence-Aware Escalation in Enterprise AI Governance: A Technical Framework

Janardhana Naidu Kola

Independent Researcher, United States of America

Abstract: The deployment of artificial intelligence systems in enterprise risk management presents a novel governance challenge: determining which AI-generated risk assessments carry sufficient confidence for automated action and which warrant human review. This is the confidence-conditioned escalation problem, where uncertainty quantification governs the boundary between automated AI and human intervention. Current enterprise governance practices rely on categorical automation rules that conflate risk category classification with model confidence, creating a critical epistemic gap. This paper formalizes the Confidence-Aware Escalation (CAE) framework, which integrates quantified uncertainty into escalation decisions via conformal prediction theory. Conformal prediction provides distribution-free, finite-sample coverage guarantees, enabling governance policies to be specified in terms of verifiable error rate bounds rather than heuristic confidence thresholds. The CAE framework classifies AI outputs into three automation tiers by jointly evaluating prediction-set cardinality and risk category. A three-layer governance architecture comprising inference production, human oversight, and regulatory compliance is proposed, supported by structured stakeholder roles: Risk Owners, Model Stewards, and Executives. Adaptive feedback learning pipelines maintain coverage guarantees under concept drift without reward-hacking incentives. A policy-driven fairness monitoring protocol resolves mathematical incompatibility among fairness criteria through organizational policy specification rather than technical compromise. Empirical validation on a publicly available financial risk benchmark dataset confirms that the framework achieves high automation rates while maintaining near-theoretical coverage guarantees and enabling transparent governance of false escalation risk. The framework is aligned with EU AI Act high-risk classification requirements and US Federal Reserve SR 11-7 model risk management guidance.

Keywords: Confidence-Aware Escalation, Human-In-The-Loop Governance, Conformal Prediction, Uncertainty Quantification, Enterprise Risk Management, Model Calibration, Fairness Monitoring, Regulatory Compliance, Automation Tiers, Adaptive Learning, Distribution Shift, Dynamic Reliability, Bias Detection.

1. Introduction

The use of artificial intelligence systems for enterprise risk management, forecasting, and compliance creates a novel governance problem not well addressed within existing corporate governance frameworks. Organizations must determine, in real-time and at scale, which AI-generated risk assessments have sufficient confidence to allow automated action and which require human review before a consequential decision is made [1]. This issue overlaps with the confidence-conditioned escalation problem, where uncertainty quantification sets the boundary between automated AI and human intervention.

Enterprise governance frameworks frequently associate categorical automation rules that combine risk category classification with model confidence. A high-confidence prediction from a well-calibrated model in a regulated risk category may be less concerning than a low-confidence prediction from a poorly calibrated model in an apparently automatable category [1]. A critical epistemic gap remains: the absence of confidence-conditioned governance in enterprise AI architectures.

This paper formalizes a mathematical framework for incorporating quantified uncertainty into escalation decisions alongside categorical risk assessment. Conformal prediction theory provides distribution-free coverage guarantees, enabling governance policies to be specified in terms of verifiable error bounds [2]. The four principal contributions are: (1) a confidence-aware escalation mechanism with provable distribution-free coverage guarantees

under split conformal prediction; (2) adaptive feedback learning pipelines using adaptive conformal prediction for concept-drift-robust threshold recalibration without reward-hacking incentives; (3) a policy-driven bias and fairness monitoring protocol that resolves fairness criterion mathematical incompatibility through organizational policy specification; and (4) empirical validation on the FICO HELOC benchmark dataset demonstrating 82.8% automation rate at alpha = 0.05 with a coverage gap of 0.0026, alongside a quantified demonstration of the Chouldechova impossibility result across risk-segment proxy groups [12][17].

2. Materials and Methods

2.1 Human-in-the-Loop AI and Interactive Machine Learning

Human-in-the-loop research provides design principles for integrating human feedback into machine learning pipelines. Active learning frameworks automatically select data points to present to human labelers to maximize label value [3]. Interactive machine learning systems face five principal challenges: operator mental models of the AI system, annotation burden, active feature space exploration, model update management, and evaluation of interactive systems [4]. The governance architecture in this paper addresses the challenge of operator mental model management by making system uncertainty visible and actionable, enabling human operators to avoid working with an incomplete understanding of the AI system's capabilities. The human factors literature has established cognitive foundations for the operator-AI interface and taxonomies of automation ranging from fully manual control to fully autonomous systems [5]. Enterprise AI systems risk automation surprise when probabilistic predictions are not accompanied by uncertainty estimates [5].

2.2 Algorithm Aversion, Appreciation, and Human-AI Calibration

Empirical research into human-algorithm interaction has uncovered biases critical to governance system design. Humans exposed to algorithm mistakes tend to underweight algorithmic predictions relative to objective accuracy (algorithm aversion) [6]. Complementary algorithm appreciation effects cause people to rely on algorithmic recommendations more than warranted [7]. Both aversion and appreciation represent miscalibrated trust failures. Communicating uncertainty clearly and creating feedback loops calibrated to supervisor mental models helps human operators relate actions to consequences [4][6].

2.3 Uncertainty Quantification: Calibration and Distribution-Free Guarantees

Calibration is a hard form of uncertainty quantification that produces an estimate of the probability that a prediction is in error. Most enterprise AI systems do not satisfy this requirement. Uncertainty quantification for deep neural networks can be provided by Monte Carlo dropout [8] or ensemble disagreement [9] approaches, both offering practically useful uncertainty measurements without explicit Bayesian models. Conformal prediction provides the most principled setting for enterprise governance, offering valid distribution-free, finite-sample coverage guarantees: for any user-specified target error rate alpha, the conformal prediction set contains the true label with probability at least 1 - alpha, regardless of the data distribution [2]. This enables the translation of enterprise risk tolerance statements into verifiably valid uncertainty specifications.

Table 1: Model Calibration Methods and Post-Hoc Recalibration [1]

Calibration Method	Technical Mechanism	Calibration Target	Application Stage	Confidence Quality
Platt Scaling	Logistic function fitting to probability scores	Binary classification calibration	Post-training score transformation	Improved probability accuracy
Isotonic Regression	Non-parametric monotonic function fitting	Multi-class calibration in ordered spaces	Post-training score recalibration	Flexible calibration without distributional assumptions

Conformal Prediction Wrapper	Nonconformity score calibration set construction	Distribution-free coverage guarantee generation	Post-training prediction set construction	Distribution-free coverage with theoretical guarantees

2.4 Regulatory Frameworks and Model Risk Management

US Federal Reserve guidance (SR 11-7) established a baseline model risk management framework requiring independent model validation, documentation of model shortcomings, and senior management oversight [10]. The EU AI Act defines four risk categories: (1) unacceptable risk (prohibited); (2) high risk (mandatory conformity assessment); (3) limited risk (transparency obligations); and (4) minimal risk. Enterprise AI systems for credit scoring, employment, critical infrastructure, and healthcare are classified as high-risk AI systems subject to human oversight, technical robustness, accuracy, transparency, and cybersecurity requirements [11].

2.5 Fairness, Bias, and Criterion Incompatibility

Fairness monitoring is technically difficult because fairness metrics are not generally mathematically compatible. When base rates of protected groups differ, no classifier can simultaneously satisfy calibration, false positive rate equality, and false negative rate equality [12]. Analogous impossibility results hold for broader classes of fairness criteria [13]. Available techniques include post-processing, adversarial debiasing, and individual fairness approaches; the appropriate choice depends on organizational context and selected fairness criterion [14].

2.6 Dataset: FICO HELOC Benchmark

This study applies the CAE framework to the FICO Home Equity Line of Credit (HELOC) dataset [17], a publicly available benchmark containing N = 10,459 applicant records across 23 financial features, widely used in explainable AI and risk governance research [18][19]. The binary outcome (RiskPerformance: Bad vs. Good) represents whether an applicant defaulted within 24 months. The dataset exhibits a near-balanced class distribution: 4,580 Bad outcomes (45.8%) and 5,879 Good outcomes (54.2%). Data are partitioned into three disjoint sets using a 60/20/20 stratified split: training set (n = 6,275), calibration set (n = 2,092), and test set (n = 2,092). Stratified splitting preserves the 45.8% bad-rate across all partitions. This strict partitioning is essential for the theoretical validity of split conformal prediction: the calibration set must be exchangeable with the test set and disjoint from the training set [16].

3. Results

3.1 Three-Layer Governance Framework Architecture

The governance architecture operationalizes confidence-conditioned governance as three structural layers, making uncertainty quantification a first-class governance input alongside risk category classification [1]. The Inference Layer produces risk estimates using classification models selected for both calibration and discriminative power. Platt scaling, isotonic regression, and conformal prediction wrappers provide calibrated prediction sets with distribution-free coverage guarantees [2]. The Human Oversight Layer comprises Risk Owners (domain experts confirming assessments), Model Stewards (technically trained governance staff), and Executives (defining enterprise policy) [1]. The Compliance Control Layer defines organizational and regulatory accountability through an automation tier framework: Tier 1 (automation permitted) for approved risk categories with low uncertainty; Tier 2 (human review required) when uncertainty exceeds thresholds or categories require review; and Tier 3 (human decision mandatory) for high-impact risk categories regardless of uncertainty [1].

Table 2: Three-Layer Governance Framework Architecture [1]

Framework Layer	**Component Function**	**Primary Responsibility**	**Accountable Stakeholder**
Inference Layer	Probability calibration and conformal prediction set generation	Risk scoring and uncertainty quantification	Model Stewards
Human Oversight Layer	Risk Owner decision structuring and Model Steward governance	Tier-conditional human review and feedback collection	Risk Owners, Model Stewards, Executives

Compliance Control Layer	Automation tier enforcement and audit trail logging	Regulatory alignment and accountability	Executives, Compliance Officers

3.2 Formal Specification of Confidence-Aware Escalation

The confidence-aware escalation decision rule maps conformal prediction sets and risk categories to automation levels. Conformal prediction sets are generated from split conformal methods on calibration datasets disjoint from training datasets, providing coverage guarantees P(y in Phi(x; alpha)) >= 1 - alpha [2]. Tier assignment rules are: Tier 1 if all risk categories belong to automation sets and prediction set cardinality falls below low-uncertainty thresholds; Tier 2 if any approved risk category exceeds low-uncertainty thresholds or review-required risk categories are present; and Tier 3 if uncertainty estimates exceed high-uncertainty thresholds or mandatory-decision risk categories are present [1]. The key formal property of CAE is that it inherits the distribution-free coverage guarantees of conformal prediction: for any alpha specified in governance policy, the probability of Tier 1 assignment conditional on true risk severity crossing the relevant action threshold is bounded by alpha regardless of model architecture or data distribution.

3.3 Adaptive Feedback Learning and Concept Drift Management

The adaptive feedback learning pipeline adapts escalation thresholds using structured Risk Owner feedback and online learning processes that avoid reward-hacking incentives [1]. The pipeline operates through three stages: (i) Feedback collection, in which Risk Owners provide correct labels and outcome reasoning; (ii) Recalibration, in which calibration datasets are updated with new labeled examples while meeting adaptive conformal prediction coverage guarantees; and (iii) Retraining, triggered when feedback volume exceeds a threshold or system performance drops [1]. Concept drift detection monitors for distributional changes indicating calibration decay using the Page-Hinkley test and other statistical tests on performance metrics [15]. Adaptive conformal prediction preserves coverage under distribution shift while avoiding reward-specification difficulties [1].

Table 3: Regulatory Alignment Mechanisms: EU AI Act and SR 11-7 [1][10][11]

HITL-G Mechanism	EU AI Act Requirement	Regulatory Article	SR 11-7 Requirement	Compliance Function
Confidence-Aware Escalation	Risk management system and human oversight	Articles 9, 14	Model output validation and action thresholds	Decision validation and operator intervention
Conformal Prediction Coverage	Accuracy, robustness, and cybersecurity standards	Article 15	Conceptual soundness and mathematical model validation	Verifiable reliability specification and error bounding
Adaptive Feedback Learning	Post-deployment testing and performance monitoring	Article 9(6)	Ongoing monitoring and degradation trigger mechanisms	Continuous performance evaluation and recalibration
Comprehensive Audit Trails	Record-keeping and traceability obligations	Article 12	Compliance evidence and auditor accessibility	Immutable append-only logs for independent auditors
Model Steward Independence	Independent review requirement	Article 9	Independent validation,	Conflict-of-interest elimination and

	before deployment		separation from development	governance oversight

3.4 Base Model Performance on HELOC Dataset

A Gradient Boosted Machine (GBM) was trained on the HELOC training set with hyperparameters: 300 trees, learning rate 0.05, maximum depth 4, subsample ratio 0.8, and minimum leaf size 30. The model achieves AUC = 0.696 on the held-out test set (Brier score = 0.221), consistent with published HELOC benchmarks [18][21]. The calibration curve shows moderate calibration with some overconfidence at the upper tail, reinforcing the case for distribution-free conformal prediction over confidence thresholds alone: the coverage guarantee does not depend on model calibration [20].

Table 4: Model Performance Summary

Metric	Training Set	Calibration Set	Test Set
N (records)	6,275	2,092	2,092
AUC	—	—	0.696
Brier Score	—	—	0.221
Bad Rate	45.8%	45.8%	46.7%

3.5 Conformal Prediction Wrapper and Escalation Tier Assignment

Split conformal prediction was applied using nonconformity scores defined as s_i = 1 - P(y_true | x_i). For binary classification, if the true label is Bad (y=1), the score is 1 - P(Bad|x); if Good (y=0), the score is P(Bad|x). The conformal threshold is computed as the ceiling[(n_cal + 1)(1 - alpha)] / n_cal quantile of calibration-set nonconformity scores. A prediction set for test point x includes class k if and only if the nonconformity score for class k does not exceed this threshold. Escalation tiers are assigned as: Tier 1 (Automated Decision) when the prediction set contains exactly one class and confidence margin exceeds 0.10; Tier 2 (Human Review Required) when the prediction set contains both classes or the single prediction lies near the decision boundary (margin <= 0.10); and Tier 3 (Mandatory Human Decision) when the prediction set is empty, indicating an anomalous input.

3.6 Empirical Results: Coverage and Automation Rate

Table 5 presents conformal prediction tier allocation results across alpha levels on the held-out test set (N_test = 2,092). At alpha = 0.05, empirical coverage is 0.9474 versus theoretical 0.9500, a gap of merely 0.0026. The slight shortfall is expected under finite samples and confirms the distribution-free coverage guarantee [20]. At alpha = 0.05, 82.8% of test cases qualify for Tier 1 automated decisions, with the remaining 17.2% requiring Tier 2 human review. No cases fall into Tier 3, indicating the HELOC distribution is well-behaved relative to the calibration set. The false escalation rate (FER) ranges from 0.223 to 0.312 across alpha levels, representing the fraction of automated decisions where the true label falls outside the single-class prediction set. Governance policy can be set directly in terms of observable error rates: a Chief Risk Officer who tolerates no more than 5% incorrect automated decisions sets alpha = 0.05, which the framework guarantees with distribution-free validity.

Table 5: Conformal Prediction Tier Allocation Results (N_test = 2,092)

a	t^	AvgSet	Tier1%	Tier2%	Tier3%	EmpCov	ThCov	CovGap	FER
0.05	0.81	1.17	82.8%	17.2%	0.0%	0.9474	0.9500	0.0026	0.223
0.10	0.74	1.09	89.4%	10.6%	0.0%	0.9008	0.9000	0.0008	0.312
0.20	0.5931	1.2868	82.8%	17.2%	0.0%	0.7777	0.8000	−0.0223	0.3117

3.7 Empirical Demonstration of Fairness Incompatibility

Test cases were partitioned using the ExternalRiskEstimate (ERE) feature as a demographic-adjacent proxy, split at the median (ERE = 66.8): High-ERE Group (ERE >= 67, n = 1,046) and Low-ERE Group (ERE < 67, n = 1,046). Results confirm the Chouldechova impossibility theorem empirically: (1) base rates differ significantly (High-ERE bad rate = 38.6%; Low-ERE bad rate = 53.0%; difference = 14.3 percentage points); (2) Statistical Parity is violated (positive prediction rate gap = 27.6 pp: 26.7% vs. 54.3%); (3) Equalized Odds is also violated (TPR gap = 26.0 pp; FPR gap = 22.9 pp); and (4) closing the SP gap would simultaneously worsen the equalized odds gap. No classifier can simultaneously satisfy SP and EO when base rates differ [12].

Table 6: Group-Level Performance and Fairness Metrics (PP = positive prediction rate; SP = statistical parity; EO = equalized odds)

Group	N	Bad Rate	AUC	PPR	TPR	FPR	Brier
High-ERE (ERE >= 67)	1,046	38.6%	0.677	26.7%	34.4%	21.5%	0.216
Low-ERE (ERE < 67)	1,046	53.0%	0.685	54.3%	60.4%	44.4%	0.225
Gap (Low - High)	—	14.3pp	0.008	27.6pp	26.0pp	22.9pp	0.009

4. Discussion

4.1 Fairness Monitoring and Policy-Driven Criterion Selection

The bias and fairness monitoring protocol implements policy-driven criterion selection, resolving incompatibility through organizational policy specification rather than technical tradeoffs [1]. Executives select from a predefined set of fairness criteria including statistical parity, equalized odds, calibration within groups, or individual fairness, and document their reasoning, monitored protected attributes, and violation tolerance thresholds as an auditable fairness policy record [1]. Metric computation runs continuously against inference layer outputs across multiple granularity levels. In the HELOC empirical context, the framework recommends Equalized Odds for three policy reasons: (i) SR 11-7 guidance requires documentation of model limitations and performance by segment; (ii) clients with similar actual cost risk should receive consistent escalation tier assignments; and (iii) false negatives carry greater financial and regulatory risk than false positives. The selected criterion is recorded in the governance audit trail with responsible executive authorization, satisfying EU AI Act Article 10 and SR 11-7 Section 5 documentation requirements.

Table 7: Calibration Quality Maintenance and Performance Monitoring [1]

Monitoring Dimension	Method	Trigger Condition	Response Action
Calibration Drift	Page-Hinkley test on rolling ECE	Statistically significant increase in calibration error	Recalibration using updated calibration set
Coverage Gap	Empirical coverage vs. theoretical (1-alpha)	Coverage gap exceeds 0.01	Recompute conformal threshold on refreshed data
Fairness Metric Drift	Statistical significance test on group-level metrics	Metric crosses tolerance threshold	Executive alert and policy review

Risk Owner Calibration	Outcome-weighted feedback scoring	Low correlation between RO decisions and outcomes	Reduce RO feedback weight; initiate training review

4.2 Regulatory Alignment and Compliance by Design

The mapping of framework mechanisms to regulatory provisions illustrates compliance-by-design properties. The Confidence-Aware Escalation mechanism meets requirements for a risk management system and human oversight by defining thresholds for action and operator involvement [1][11]. Conformal prediction coverage guarantees quantify limitations on model accuracy, robustness, and cybersecurity [1][11]. Audit trails capture AI output values, uncertainty measures, classification tiers, review decisions, rationales, evaluation metrics, and timestamps in append-only immutable logs accessible to independent compliance auditors [1][10][11]. Adaptive feedback learning pipelines track performance over time and intervene upon drift, satisfying monitoring and post-deployment testing requirements. Model Steward independent validation decouples validation from development, operationalizing independent review requirements [1][10].

4.3 Organizational Structure and Implementation Challenges

Model Stewards should be functionally independent from model development teams to create appropriate accountability and align organizational incentives [1]. Executive policy accountability for fairness criterion selection, combined with tracking of Risk Owner decision quality against real-world outcomes, provides accountability lacking in conventional human review systems [1]. Risk Owner feedback is weighted by calibration data correlated with future outcomes, reducing the influence of poorly calibrated Risk Owners and maximizing the value of human domain knowledge [1]. Implementation challenges include: (1) conformal prediction coverage guarantees rely on exchangeability assumptions; (2) distributional drift detection alone does not maintain coverage properties during adaptation; (3) adaptive feedback learning is limited by the frequency and quality of outcome tracking; and (4) fairness monitoring statistical tests lack power in small protected group populations.

5. Conclusion

The Human-in-the-Loop AI Governance framework addresses an epistemological gap in enterprise AI governance by making quantified uncertainty a first-class governance input alongside categorical risk classification. The Confidence-Aware Escalation mechanism's formally specified decision logic, rooted in conformal prediction theory, enables organizations to specify governance policies in terms of verifiable error rate bounds rather than heuristic confidence thresholds. The four principal contributions are: (1) confidence-aware escalation mechanisms with provable coverage guarantees under split conformal prediction; (2) adaptive feedback learning pipelines for concept-drift-robust threshold recalibration; (3) policy-driven bias and fairness monitoring protocols that resolve criterion mathematical incompatibility; and (4) regulatory alignment evidence demonstrating compliance-by-design properties under the EU AI Act and US Federal Reserve SR 11-7. Future work requires empirical studies of calibration quality across diverse risk domains, controlled experiments on conformal prediction uncertainty communication versus conventional confidence scores, and longitudinal governance quality assessment of deployed systems. Responsible AI governance requires technical architecture to be complemented by governance-enabling organizational structures: Model Stewards independent of development teams, executives accountable for fairness criterion selection, and systematic outcome tracking.

References

1. I. W. Eisenberg et al., "The Unified Control Framework: Establishing a Common Foundation for Enterprise AI Governance, Risk Management, and Regulatory Compliance," arXiv, 2025. Available at: https://arxiv.org/pdf/2503.05937 (Accessed on date: 23 June 2026).
2. T. Kaiser and P. Herzog, "A Tutorial on Distribution-Free Uncertainty Quantification Using Conformal Prediction," Association for Psychological Science, 2025. Available at: https://journals.sagepub.com/doi/pdf/10.1177/25152459251380452 (Accessed on date: 23 June 2026).
3. B. Settles, "Active Learning Literature Survey," Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009. Available at: https://burrsettles.com/pub/settles.activelearning.pdf (Accessed on date: 23 June 2026).
4. G. C. Saha et al., "Human-AI Collaboration: Exploring Interfaces for Interactive Machine Learning," Tuijin Jishu/Journal of Propulsion Technology, 2023.
5. R. Parasuraman et al., "A model for types and levels of human interaction with automation," IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, vol. 30, no. 3, pp. 286-297, 2000.

6. E. Jussupow et al., "Why Are We Averse to Algorithms? A Comprehensive Literature Review On Algorithm Aversion," AIS Electronic Library (AISeL), 2020.
7. J. M. Logg et al., "Algorithm appreciation: People prefer algorithmic to human judgment," Organizational Behavior and Human Decision Processes, vol. 151, pp. 90-103, 2019.
8. Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proceedings of ICML, pp. 1050-1059, 2016.
9. B. Lakshminarayanan et al., "Simple and scalable predictive uncertainty estimation using deep ensembles," in Advances in Neural Information Processing Systems, vol. 30, pp. 6402-6413, 2017.
10. Board of Governors of the Federal Reserve System, "Supervisory Guidance on Model Risk Management," SR Letter 11-7, April 4, 2011. Available at: https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm (Accessed on date: 23 June 2026).
11. M. Veale and F. Z. Borgesius, "Demystifying the Draft EU Artificial Intelligence Act," Computer Law Review International, arXiv, 2021. Available at: https://arxiv.org/pdf/2107.03721 (Accessed on date: 23 June 2026).
12. A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," arXiv, 2017. Available at: https://arxiv.org/pdf/1703.00056 (Accessed on date: 23 June 2026).
13. J. Kleinberg et al., "Inherent trade-offs in the fair determination of risk scores," arXiv, 2016. Available at: https://arxiv.org/pdf/1609.05807 (Accessed on date: 23 June 2026).
14. B. H. Zhang et al., "Mitigating unwanted biases with adversarial learning," ACM Digital Library, 2018.
15. J. Gama et al., "A survey on concept drift adaptation," ACM Digital Library, 2014.
16. A. N. Angelopoulos and S. Bates, "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification," arXiv:2107.07511, 2021. Available at: https://arxiv.org/pdf/2107.07511 (Accessed on date: 23 June 2026).
17. FICO, "Explainable Machine Learning Challenge," Fair Isaac Corporation, 2018. Available at: https://community.fico.com/s/explainable-machine-learning-challenge (Accessed on date: 23 June 2026).
18. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, pp. 206-215, 2019.
19. Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," in Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 150-158, 2012.
20. V. Vovk, A. Gammerman, and G. Shafer, Algorithmic Learning in a Random World. Springer, 2005.
21. M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in Proceedings of the AAAI Conference on Artificial Intelligence, pp. 1527-1535, 2018.