Trustworthy Abnormality Detection from Welding Images Through Class-Conditional Conformal Learning and Bayesian Cost Minimization

Zhenling Chen[1], Zhiguo Zeng[2]

[1] *Laboratoire Genie Industriel, Centralesupelec, Universite Paris-Saclay*
zhenling.chen@student-cs.fr

[2] *Chair on Risk and Resilience of Complex Systems,*
Laboratoire Genie Industriel, Centralesupelec, Universite Paris-Saclay
zhiguo.zeng@centralesupelec.fr

ABSTRACT

Industrial fault/abnormality detection is often criticized for lacking explainability and robustness. Furthermore, practical industrial datasets are frequently highly imbalanced and operate under extreme risk asymmetry, *i.e.*, false negatives carry penalties orders of magnitude higher than false alarms, which poses significant challenges to reliable detection. In this paper, we develop a trustworthy AI framework to improve confidence in welding defect detection. The proposed framework integrates two primary techniques: class-conditional conformal learning and Bayesian cost minimization. First, the conformal learning model quantifies the trustworthiness of model predictions. Instead of forcing a binary classification, the model outputs an "uncertain" state when confidence is low, facilitating informed human intervention. Second, a Bayesian cost minimization algorithm is used to avoid over-conservative predictions that yield too many "uncertain" predictions. Results on a real-world welding quality inspection dataset show that the developed method adapts robustly to dynamic intervention costs and mitigates worst-case cost spikes. The framework is deployment-oriented: it is not uniformly optimal in every setting, but it consistently avoids catastrophic failures and maintains a favorable cost–accuracy–intervention trade-off across heterogeneous base-model qualities.

1. INTRODUCTION

Safety-critical industrial manufacturing relies on automated abnormality detection systems that must be both highly accurate and economically robust. In practical deployment, these operational environments are frequently characterized by severe class imbalance and extreme risk asymmetry (ETAI Association, n.d.): the penalty for a missed defect (false negative) often outweighs a false alarm (false positive) by orders of magnitude, such that $C_{FN} \gg C_{FP}$.

Under such strict constraints, traditional decision-making relies heavily on Empirical Risk Minimization (ERM) driven by Bayes risk (Elkan, 2001). However, Bayes-risk decisions are notoriously fragile. Deep neural networks frequently exhibit severe overconfidence, particularly on rare fault classes. Any slight probability distortion—often exacerbated by global post-hoc calibration methods that optimize for aggregate metrics rather than minority preservation—can trigger catastrophic missed detections and severe downstream financial losses.

To bypass the reliance on perfectly calibrated heuristics, Conformal Prediction (CP) (Vovk, Gammerman, & Shafer, 2005) offers a rigorous framework for uncertainty quantification. By outputting set-valued predictions $\Gamma^{\alpha}(x)$, CP allows the model to yield an "uncertain" state, facilitating informed human intervention. Recent advancements have extended CP to handle class imbalance via Class-Conditional Conformal Prediction (CCCP) (Sadinle, Lei, & Wasserman, 2019; Romano, Sesia, & Candès, 2020). Yet, pure CP methodologies suffer from inherent *cost-blindness* (Stankevičiūtė & et al., 2021; Bates, Angelopoulos, Lei, Malik, & Jordan, 2021). They construct statistical boundaries based purely on non-conformity, ignoring the financial penalty of human review (C_{UNK}). When intervention costs scale up dynamically, pure CP systems output financially unsustainable levels of ambiguity.

To bridge the gap between reliable defect isolation and operational cost, we propose a trustworthy AI framework for generic abnormality detection, which we validate on a real-world welding inspection dataset. Our approach introduces a

Code availability: https://github.com/JialingRichard/CCCP-Bayes

decoupled fail-safe cascade: first, a Class-Conditional Conformal Prediction (CCCP) stage isolates uncertain predictions, ensuring robust coverage even under extreme class imbalance. Second, a Bayes-risk decision maker performs economic cost minimization exclusively within this isolated region to filter out overly conservative predictions.

The rest of this paper is organized as follows. Sect. 2 reviews related work and clarifies the research gap. Sect. 3 formulates the problem and presents the developed method. Sect. 4 presents the welding fault detection dataset and experiment set-ups. Sect. 5 presents the results of the experiment. Sect. 6 presents ablation studies to better understand the performance of the developed method. Sect. 7 concludes the paper.

2. RELATED WORK AND RESEARCH GAP

The action-space model in this paper follows classical cost-sensitive decision making with a reject/intervention option (Elkan, 2001; Chow, 1970) and selective classification (Geifman & El-Yaniv, 2017). The research gap lies in how to instantiate this known decision structure for welding inspection, where rare defects, severe posterior miscalibration, class-conditional coverage failure, and dynamically scaled intervention costs must be handled together. Existing strategies usually address only one side of this problem: Bayes-risk rules minimize expected cost but depend strongly on posterior quality, whereas CP/CCCP isolate uncertainty but ignore the cost of intervention.

Value-of-Information (VoI) is a natural Bayesian alternative for deciding whether additional information should be acquired (Howard, 1966). In the official OK/KO/UNKNOWN protocol studied here, however, UNKNOWN is a fixed-cost terminal intervention that returns a reliable label and has no additional downstream decision stage. Under this assumption, VoI is exactly the same comparison as the Bayes-risk comparison with the UNK action: choose UNKNOWN precisely when its expected intervention risk is lower than the risks of declaring OK or KO. Therefore, VoI is not a separate missing baseline in the present setting; it is implemented by the UNK branch of the Bayesian cost minimization step. Non-trivial VoI becomes distinct only when inspection is noisy, sequential, time-varying, or coupled to later decisions.

Risk-Controlling Prediction Sets (RCPS) (Bates et al., 2021) are also relevant because they aim to control user-specified expected risks. Under the extreme asymmetric cost matrix considered here, however, feasibility can be violated and the resulting policy can become dominated by abstention. We therefore report RCPS later as an exploratory limitation case rather than as a core baseline.

3. METHODOLOGY

3.1. Problem Formulation

Following cost-sensitive decision making with reject and intervention options (Elkan, 2001; Chow, 1970; Geifman & El-Yaniv, 2017), let $x \in \mathcal{X}$ denote input features and $y \in \mathcal{Y} = \{0, 1\}$ be the true label, where 0 represents normal operations (*ok*) and 1 denotes a fault state (*ko*). Unlike conventional binary classification, a trustworthy fault detection system must select an action from three possible states $a \in \mathcal{A} = \{0, 1, \text{UNK}\}$, where UNK means that the algorithm is unsure about the true state and human intervention is needed.

Let $\mathcal{C} : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}^+$ denote a cost matrix that maps the model prediction to a reward/penalty by comparing it to the ground truth. In this paper, we assume an extreme asymmetrical risk matrix: the false negative penalty strictly dominates the false positive penalty ($C_{FN} \gg C_{FP}$). This reflects the reality of fault detection in safety-critical systems, where the risk of missing an alarm is much higher than that of reporting a false alarm. When the model chooses to report UNK, an intervention cost of C_{UNK} is charged because the field operator needs to intervene. The objective of trustworthy fault detection is to formulate a decision rule $\pi : \mathcal{X} \to \mathcal{A}$ that minimizes the expected cost $\mathbb{E}[\mathcal{C}(Y, \pi(X))]$ while reducing maximum regret over all possible scenarios.

3.2. Limitations of Existing Strategies

In the literature, there are two main methods used for trustworthy fault detection. The first is empirical risk minimization based on Bayes-risk decision rules. In this approach, a machine-learning model is trained to output a posterior probability distribution $\hat{p}(y \mid x)$. The model then chooses the best prediction based on the average posterior costs:

$$a^*(x) = \arg\min_{a \in \mathcal{A}} \sum_{y \in \mathcal{Y}} \hat{p}(y \mid x) \mathcal{C}(y, a). \tag{1}$$

This approach is theoretically optimal only if $\hat{p}(y \mid x)$ is perfectly calibrated. However, real fault detection data are usually highly imbalanced, which can bias the estimated $\hat{p}(y \mid x)$ and lead to sub-optimal or even completely wrong decisions.

Conversely, Conformal Prediction (CP) constructs a set-valued prediction $\Gamma^\alpha(x) \subseteq \mathcal{Y}$ that guarantees true label coverage $P(y \in \Gamma^\alpha(x)) \geq 1 - \alpha$. While mathematically elegant, pure CP architectures do not account for class-dependent label distributions. One drawback is that, when faced with high uncertainty, CP tends to include multiple labels in the prediction sets $\Gamma^\alpha(x)$, (e.g., $\{0, 1\}$), forcing the system to return many UNK actions. As C_{UNK} scales up, the expected system cost increases dramatically, rendering pure CP financially unsustainable.

3.3. The Mixed CCCP-BCM Framework

To mitigate the fragility of uncalibrated probabilities and the cost-blindness of CCCP and Bayesian Cost Minimization (BCM), respectively, we propose a framework that mixes the two approaches. The framework first leverages CCCP to screen all possible uncertain predictions. More specifically, we utilize the inverse probability as the heuristic non-conformity score: $s(x, y) = 1 - \hat{p}(y \mid x)$. Using a calibration set $\mathcal{D}_{cal}$, we partition it by class labels $\mathcal{D}^y_{cal}$ and compute the class-conditional empirical quantile:

$$\begin{aligned} &\hat{q}_y = \text{the } k_y\text{-th smallest value in } \{s(x_i, y)\}_{i \in \mathcal{D}^y_{cal}}, \\ &k_y = \lceil (n_y + 1)(1 - \alpha) \rceil. \end{aligned} \tag{2}$$

To ensure deployment simplicity and parameter parity, we enforce a strict single-tuning constraint $\alpha_0 = \alpha_1 = \alpha$. The CCCP filter generates a valid prediction set:

$$\Gamma^\alpha(x) = \{y \in \mathcal{Y} \mid s(x, y) \leq \hat{q}_y\}. \tag{3}$$

It should be noted that although $\alpha_0 = \alpha_1$, the threshold quantiles $\hat{q}_y$ differ between $y = 0$ and $y = 1$, allowing us to capture the class-dependent distribution shifts on y.

As shown in the previous section, one major drawback of CCCP is that it tends to be over-conservative and produce too many "unknowns". To address this issue, in the second stage, we apply Bayesian cost minimization to the uncertain predictions identified from the CCCP. The purpose is to filter out those over-conservative "unknown" predictions from the CCCP. More specifically, when $\Gamma^\alpha(x)$ is empty or contains multiple labels ($|\Gamma^\alpha(x)| \neq 1$), CCCP has no unique conformal decision, and the system invokes Bayes-risk settlement via Equation 1. We do not truncate or renormalize posterior probabilities; $\Gamma^\alpha(x)$ is used only as a gating trigger region. The intervention branch is fully implementable because the UNK action uses expected intervention risk,

$$C_{\text{UNK}}(x) = \hat{p}(ok \mid x)\, C^{ok}_{UNK} + \hat{p}(ko \mid x)\, C^{ko}_{UNK}. \tag{4}$$

In implementation, this expected risk is used as the UNK action cost inside Bayes fallback evaluation. This gating design confines posterior-sensitive Bayes decisions to the ambiguity region and provides a practical fail-safe mechanism.

The developed framework is summarized in Algorithm 1.

For example, if a test image yields $\Gamma^\alpha(x) = \{ok\}$, the first stage has a unique conformal decision and the system directly outputs ok. If instead $\Gamma^\alpha(x) = \{ok, ko\}$, or if $\Gamma^\alpha(x) = \emptyset$, the conformal stage has no unique decision; the image then enters Stage 2, where the risks of declaring ok, declaring ko, and requesting UNK intervention are compared by Equation 1.

Require: x, $\hat{q}_{ok}, \hat{q}_{ko}$ from Eq. 2, $\mathcal{C}$, $\hat{p}(y \mid x)$

// Stage 1: Topological Isolation (CCCP, Eq. 3)
Initialize ambiguity set $\Gamma^\alpha(x) \leftarrow \emptyset$
for $y \in \{ok, ko\}$ **do**
if $1 - \hat{p}(y \mid x) \leq \hat{q}_y$ **then**
$\Gamma^\alpha(x) \leftarrow \Gamma^\alpha(x) \cup \{y\}$
end if
end for
// Stage 2: Economic Settlement (Bayes-Risk Gating)
if $|\Gamma^\alpha(x)| = 1$ **then**
Return $a^* \leftarrow y \in \Gamma^\alpha(x)$
else
Compute $C_{\text{UNK}}(x)$ by Eq. 4
Select a^* by Bayes risk in Eq. 1
Return a^*
end if

Algorithm 1. Mixed CCCP-BCM.

4. EXPERIMENTAL SETUP

To validate the robustness of the decoupled framework, we use the public Welding Quality Detection Challenge benchmark and its official OK/KO/UNKNOWN protocol (ETAI Association, n.d.). The objective is to detect welding quality faults from welding images. Figure 1 shows examples of OK and KO welds. One difficult issue in this dataset is that some images have poor visual quality. The fault detection algorithm needs to identify these samples as "unknown" rather than blindly guess an answer.

Figure 1. An example of the welding images.

We evaluate under an extreme asymmetric cost matrix where $C_{FN} = 3000$ and $C_{FP} = 30$, as detailed in Table 1. The official operating point is $\lambda = 1$ (baseline intervention cost). We additionally introduce $\lambda = 10$ as a stress-test setting for high-intervention deployment scenarios. The codebase and experiment artifacts are publicly available at (Chen & Zeng, n.d.).

Stress testing scales only intervention costs:

$$C^{ok}_{UNK}(\lambda) = 20\lambda, \quad C^{ko}_{UNK}(\lambda) = 41\lambda, \quad \lambda \in \{1, 10\}. \tag{5}$$

The evaluation utilizes 100 random seeds to ensure statistical

Table 1. Official challenge cost matrix ($\mathcal{C}$).

Action (a)	**True: *ok***	**True: *ko***
Normal ($a = 0$)	$C_{TN} = 0.4$	$C_{FN} = 3000.0$
Fault ($a = 1$)	$C_{FP} = 30.0$	$C_{TP} = 26.4$
Unsure ($a =$ UNK)	$C^{ok}_{UNK} = 20$	$C^{ko}_{UNK} = 41$

significance. In these 100-seed runs, backbone checkpoints and cached logits are fixed (`post-seed-only` protocol); seed randomness comes from calibration fit/select partition and alpha selection. We do not retrain the backbone per seed. Therefore deterministic policies (e.g., Baseline-2C and BayesRisk) can show near-zero across-seed variance. The CCCP mechanism is strictly constrained by a single operational parameter ($\alpha_{ok} = \alpha_{ko}$).

4.1. Dataset Provenance and Test Construction

The data source is the public Welding Quality Detection Challenge benchmark (ETAI Association, n.d.). We follow a fixed split manifest (`train/cal/test`) throughout all experiments to avoid split-induced variance across methods. The split is highly imbalanced by label: train has 15,579 *ok* vs. 345 *ko*, calibration has 2,228 *ok* vs. 49 *ko*, and test has 4,452 *ok* vs. 100 *ko*, corresponding to an overall *ok*:*ko* ratio of approximately 45:1.

4.2. CP/CCCP α Search Protocol

The conformal level α is a hyper-parameter that needs tuning. The tuning is performed only on calibration data and never on test data. For each seed, we first randomly partition calibration paths into a fit subset and a selection subset with ratio `cal_fit_frac`=0.5. The fit subset is used to estimate CP/CCCP non-conformity quantiles, and the selection subset is used to evaluate downstream cost.

4.3. Base Model Capability Profile

To make the post-processing comparison interpretable, we explicitly report base-model capability profiles before any conformal or Bayesian risk reduction. The baseline performance is obtained by using ResNet-18 as the base model and fine-tuning it on the training dataset. We test two strategies: direct fine-tuning (Vanilla-E8) and class-weighted fine-tuning (Weighted-E8). The second is introduced because the dataset is highly imbalanced. Table 2 shows a clear tension between discrimination and calibration: weighted training substantially improves minority detection quality (higher Recall$_{ko}$/AUPRC), while the same models exhibit worse probability calibration (higher ECE/NLL).

5. MAIN RESULTS

We evaluate the framework's ability to minimize expected costs while reducing high-regret outcomes across varying C_{UNK} scales.

Table 2. Base-model capability profile on representative E8 bases.

Base model	Recall$_{ko}$ (TPR$_{ko}$)	AUPRC	ECE	NLL
ResNet-Vanilla-E8	0.4975	0.7174	0.0096	0.0484
ResNet-Weighted-E8	0.9075	0.7892	0.0234	0.0820

AUPRC: Area under precision-recall curve
TPR: True Positive Rate
ECE: Expected Calibration Error
NLL: Negative Log-Likelihood

Cost-Adaptive Behavior: Fig. 2 and Table 3 show that the key deployment issue is how each method reacts when the cost of intervention changes. Pure CP/CCCP methods preserve uncertainty information, but their cost can grow sharply when UNK becomes expensive. Pure BayesRisk avoids abstention, but remains exposed to posterior errors under class imbalance. The hybrid strategies use conformal prediction only as a gate and then resolve gated samples by expected cost, which explains their more stable stress behavior. The resulting boundary is base-model dependent: CP+BayesRisk is more efficient on the stronger weighted model, whereas mixed CCCP-BCM is preferable on the weaker vanilla model under stress. This supports the claim that the two-stage method reduces cost in the evaluated stress condition, while also clarifying that it is a deployment rule rather than a uniformly dominant optimizer.

Bounded Deployment Behavior: Table 3 reports all six non-TS strategies in the core comparison using only cost-centric metrics across the official regime ($\lambda = 1$) and stress regime ($\lambda = 10$). No single strategy dominates every base condition: CP+BayesRisk is more cost-efficient on the weighted E8 base, whereas mixed CCCP-BCM is preferable on the vanilla E8 base under stress. Temperature-scaling variants are intentionally moved to the ablation section, because they are calibration interventions rather than core deployment policies.

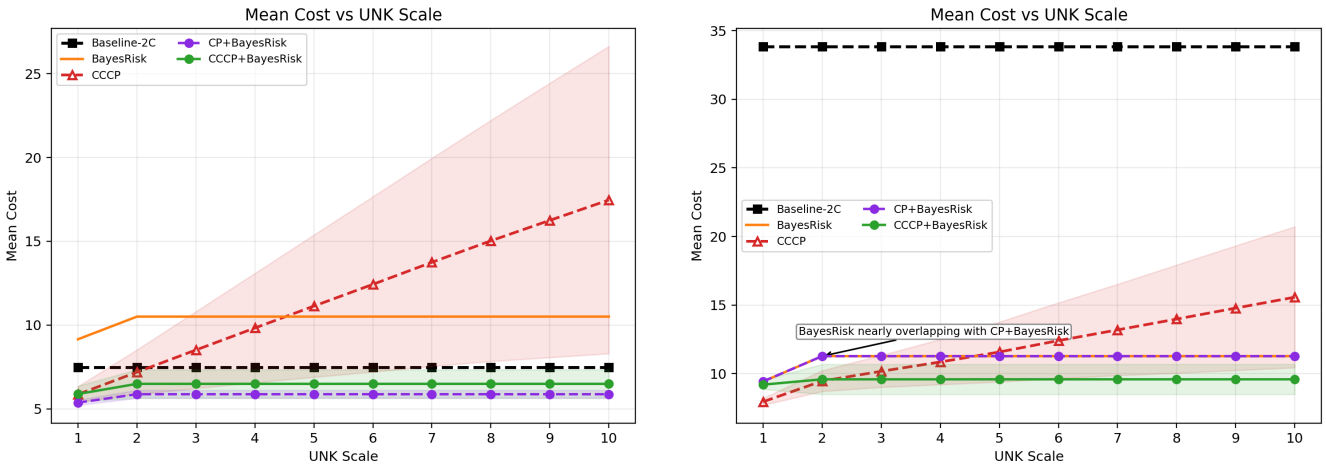
5.1. Applicability Boundary and Deployment Guidance

Under UNK$\times$10, cascade preference is base-dependent. On the weighted E8 base, both conformal filters keep high stage-1 defect recall, and CP+BayesRisk is more cost-efficient. On the vanilla E8 base, stage-1 CP recall collapses to 6.87%, while CCCP preserves 86.89% and yields lower downstream cost. Extended boundary and full 4-base robustness rankings are provided in supplementary material.

Because hybrid strategies route all non-singleton cases to Bayes fallback, stage-1 set-size rates in Table 5 are not equivalent to final UNK rates in Table 4. Under the stress setting ($\lambda = 10$), final UNK rate becomes zero for hybrid policies due to the deployed cost matrix: with $p = \hat{p}(ko \mid x)$, fallback risks are

$$R_{ok}(x) = 0.4 + 2999.6p,$$
$$R_{ko}(x) = 30 - 3.6p,$$
$$R_{\text{UNK}}(x) = 200 + 210p.$$

Hence $R_{\text{UNK}}(x) > R_{ko}(x)$ for all $p \in [0, 1]$, so UNK is never



(a) ResNet-Weighted-E8

(b) ResNet-Vanilla-E8

Figure 2. Mean cost trend under intervention-cost scaling (C_{UNK} scale from 1 to 10) for two representative base models.

Table 3. Core cost results on representative E8 bases.

Base Model	Strategy	Cost@$\lambda = 1$ (Mean $\pm$ Std)	Cost@$\lambda = 10$ (Mean $\pm$ Std)	ΔCost$(10-1)$
ResNet-Weighted-E8	Baseline-2C	7.4852 ± 0.0000	7.4852 ± 0.0000	0.0000
ResNet-Weighted-E8	BayesRisk	9.1531 ± 0.0000	10.5049 ± 0.0000	1.3518
ResNet-Weighted-E8	CP	5.2816 ± 0.1511	17.1240 ± 1.6220	11.8424
ResNet-Weighted-E8	CCCP	5.8661 ± 0.4666	17.4643 ± 9.1672	11.5982
ResNet-Weighted-E8	CP+BayesRisk	5.3735 ± 0.1539	5.8738 ± 0.2406	0.5003
ResNet-Weighted-E8	mixed CCCP-BCM (Ours)	5.8881 ± 0.4677	6.4925 ± 0.8465	0.6044
ResNet-Vanilla-E8	Baseline-2C	33.8523 ± 0.0000	33.8523 ± 0.0000	0.0000
ResNet-Vanilla-E8	BayesRisk	9.4372 ± 0.0000	11.2822 ± 0.0000	1.8450
ResNet-Vanilla-E8	CP	7.9782 ± 0.2546	23.3590 ± 0.6759	15.3808
ResNet-Vanilla-E8	CCCP	7.9592 ± 0.2496	15.5769 ± 5.1364	7.6177
ResNet-Vanilla-E8	CP+BayesRisk	9.4425 ± 0.0533	11.2822 ± 0.0000	1.8397
ResNet-Vanilla-E8	mixed CCCP-BCM (Ours)	9.2058 ± 0.3367	9.5940 ± 1.0936	0.3882

*Reading guide: under stress ($\lambda = 10$), CP+BayesRisk has the lowest mean cost on the weighted E8 base, while mixed CCCP-BCM has the lowest mean cost on the vanilla E8 base. Among Bayes-fallback hybrids (CP+BayesRisk vs mixed CCCP-BCM), mixed CCCP-BCM also has the smaller stress gap on the vanilla E8 base.

Bayes-optimal in fallback at $\lambda = 10$. Therefore zero final UNK rate in Table 4 is expected under this stress matrix, not a contradiction.

Based on this boundary behavior, a practical deployment rule is to select the cascade by calibration-time CP minority recall:

$$\begin{cases} \text{use mixed CCCP-BCM}, & \hat{r}_{ko}^{CP} < \tau, \\ \text{use CP+BayesRisk}, & \text{otherwise}, \end{cases} \tag{6}$$

where $\hat{r}_{ko}^{CP}$ is CP stage-1 defect recall on calibration data and τ is a safety threshold (e.g., $\tau = 0.2$). This turns the observed trade-off into an explicit operational policy.

6. Ablation and Discussion

6.1. Objective Mismatch of Global Calibration

A common misconception is that post-hoc probability calibration universally improves risk-based decision making. Table 6 summarizes scene-wise Wilcoxon evidence across all 8 scenarios. BayesRisk-TS vs. BayesRisk has positive mean cost difference in aggregate ($\Delta\mu = +0.6250$), and remains significant in 8/8 scenarios after Holm correction. This indicates that globally scaling logits to optimize Negative Log-Likelihood (NLL) can blur hard minority boundaries and amplify downstream C_{FN}-dominated losses. One plausible mechanism is the severe label imbalance (*ok*:*ko* ≈ 45:1): single-temperature TS is optimized for global NLL and therefore dominated by majority-class statistics. Under our highly asymmetric cost matrix, even small minority-class probability distortions can be amplified into large deployment-cost differences.

6.2. RCPS Under Extreme Asymmetric Costs

As noted in Sect. 2, we additionally evaluated RCPS (Bates et al., 2021) in exploratory runs under the same post-seed-only protocol, but on a separate setup from the main E8 comparison tables. In these trials, RCPS became difficult to use under our extreme asymmetric cost range because feasibility was often violated and the policy tended toward overly conservative

Table 4. Stress-test metrics at $\lambda = 10$ (non-TS strategies).

Base Model	Strategy	Recall$_{ko}$ @ $\lambda = 10$	UNK-Rate @ $\lambda = 10$
ResNet-Weighted-E8	Baseline-2C	0.9075	0.0000
ResNet-Weighted-E8	BayesRisk	0.9700	0.0000
ResNet-Weighted-E8	CP	0.8502	0.0638
ResNet-Weighted-E8	CCCP	0.8947	0.0630
ResNet-Weighted-E8	CP+BayesRisk	0.9582	0.0000
ResNet-Weighted-E8	mixed CCCP-BCM (Ours)	0.9576	0.0000
ResNet-Vanilla-E8	Baseline-2C	0.4975	0.0000
ResNet-Vanilla-E8	BayesRisk	0.8550	0.0000
ResNet-Vanilla-E8	CP	0.0687	0.0549
ResNet-Vanilla-E8	CCCP	0.8689	0.0393
ResNet-Vanilla-E8	CP+BayesRisk	0.8550	0.0000
ResNet-Vanilla-E8	mixed CCCP-BCM (Ours)	0.8884	0.0000

Table 5. Gating behavior of hybrid policies at $\lambda = 10$ (100-seed mean).

Base	Hybrid Policy	$P(\lvert\Gamma\rvert = 0)$	$P(\lvert\Gamma\rvert = 1)$	$P(\lvert\Gamma\rvert = 2)$	Hybrid vs Bayes disagreement
ResNet-Weighted-E8	CP+BayesRisk	0.0690	0.9310	0.0000	0.1828
ResNet-Weighted-E8	mixed CCCP-BCM	0.0666	0.9334	0.0001	0.1631
ResNet-Vanilla-E8	CP+BayesRisk	0.0719	0.9281	0.0000	0.0000
ResNet-Vanilla-E8	mixed CCCP-BCM	0.0410	0.8897	0.0693	0.0173

abstention. As a result, RCPS did not provide a competitive operating point versus BayesRisk/mixed CCCP-BCM in our exploratory setting. We therefore treat RCPS here as a limitation case rather than a core quantitative baseline in the main table set.

6.3. Why Class-Conditional Coverage is Essential

The choice of Class-Conditional CP over standard marginal CP introduces a deliberate trade-off between baseline efficiency and worst-case robustness. On the well-calibrated weighted E8 model, marginal CP+BayesRisk operates with tighter quantiles, yielding a slightly lower mean cost than our CCCP cascade. However, this empirical efficiency is highly fragile.

When deployed on the pathological vanilla E8 model, standard CP suffers from severe minority-class coverage collapse. As shown in Table 4, its defect recall drops to 6.87%. Table 5 further shows $P(|\Gamma| = 2) = 0$ and zero final-action disagreement between CP+BayesRisk and BayesRisk, which explains why CP+BayesRisk degenerates to pure uncalibrated BayesRisk (costing 11.2822).

In contrast, by enforcing conditional parity ($\alpha_{ok} = \alpha_{ko}$), our CCCP mechanism forcefully maintains a Recall$_{ko}$ of 86.89% under the exact same pathological conditions. The marginal cost increase observed on the well-calibrated base can be viewed as a necessary *fail-safe insurance premium*. This structural intervention successfully captures minority defects into the $\Gamma^{\alpha}(x)$ ambiguity set, triggering the downstream Bayes-risk settlement and pulling the expected cost down to 9.5940. Crucially, this robust fail-safe protection is achieved strictly via an $\mathcal{O}(1)$ scalar threshold comparison, incurring zero additional computational overhead.

This behavior is consistent with Table 2: BayesRisk-style decisions are more sensitive to probability calibration quality, while CCCP-style filtering is more sensitive to minority coverage structure. The proposed cascade exploits this complementarity instead of assuming either component is uniformly reliable.

Table 6. Wilcoxon summary across 8 scenarios.

Comp.	Sig.	Mean $\Delta\mu$	Mean d_z	Med. p_{holm}
Ours-BR	7/8	-1.7745	-4.1989	6.43×10^{-16}
CP+BR-BR	6/8	-1.7273	-8.8120	1.85×10^{-16}
TS-BR	8/8	+0.6250	+0.2234	3.11×10^{-11}
CCCP-CP	4/8	-0.8531	+0.0488	9.97×10^{-2}

*Note: Mean $\Delta\mu$, mean d_z, and significant-scene counts are aggregated from scene-level paired Wilcoxon test logs. Significance uses Holm-corrected p values. With 8 scenarios, median p_{holm} is computed as the average of the 4th and 5th sorted values. Abbreviations: BR = BayesRisk, TS = BayesRisk-TS.

6.4. Theoretical Limitations

Our guarantees are limited to the conformal stage. Specifically, Stage-1 CCCP provides distribution-free coverage control, but the final deployed action from the gated hybrid policy does not inherit a distribution-free decision-risk bound. In addition, Bayes fallback still depends on posterior quality $\hat{p}(y \mid x)$, so the method is not mathematically immune to posterior misspecification. Finally, our alpha-selection rule (*top-5% + median*) and "maximum regret" claims are empirical protocol choices; they are not presented as formal optimality or minimax-regret theorems.

7. CONCLUSION

This paper presents a mixed CCCP-BCM framework designed for trustworthy fault detection under extreme asymmetric risk environments. By structurally separating distribution-free ambiguity filtering from Bayesian economic settlement, the framework improves scenario-wise robustness in our evaluated settings. Extensive evaluations demonstrate that the method reliably adapts to dynamic intervention costs and provides a structural fail-safe against both uncalibrated probabilities

and minority-class degradation. Overall, the method is not uniformly optimal for every base model and cost regime, but it consistently improves worst-case behavior and preserves a favorable cost–accuracy–intervention trade-off, making it deployment-ready for safety-critical industrial screening.

The main limitation is that the distribution-free guarantee applies to the Stage-1 conformal set, whereas the final deployed action after Bayes fallback still depends on posterior quality and on the fixed reliable-intervention assumption. Future work should extend the decision layer to non-trivial VoI settings with imperfect inspection, sequential re-testing, time-varying intervention costs, and broader PHM datasets.

REFERENCES

Bates, S., Angelopoulos, A., Lei, L., Malik, J., & Jordan, M. I. (2021). Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*.

Chen, Z., & Zeng, Z. (n.d.). *Cccp-bayes (code and experiment artifacts)*. `https://github.com/JialingRichard/CCCP-Bayes`. (Accessed: 27 Feb 2026)

Chow, C. K. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, *16*(1), 41–46.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the international joint conference on artificial intelligence (ijcai)*.

ETAI Association. (n.d.). *Welding quality detection challenge (official website, dataset and evaluation protocol)*. `https://etaia.github.io/Welding-Quality-Detection-Challenge/`. (Accessed: 27 Feb 2026)

Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In *Proceedings of the neural information processing systems (neurips)*.

Howard, R. A. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, *2*(1), 22–26.

Romano, Y., Sesia, M., & Candès, E. (2020). Classification with valid and adaptive coverage. In *Proceedings of the neural information processing systems (neurips)*.

Sadinle, M., Lei, J., & Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association (JASA)*.

Stankevičiūtė, K., & et al. (2021). Conformal time-series forecasting. In *Proceedings of the neural information processing systems (neurips)*.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.

BIOGRAPHIES

Zhenling Chen is working towards his PhD degree in Laboratoire Genie Industriel, Centralesupelec, Universite Paris-Saclay, France.

Zhiguo Zeng is a professor at Laboratoire Genie Industriel, Centralesupelec, Universite Paris-Saclay, France. His research interests include reliability and resilience modeling and optimization, predictive maintenance and AI-driven reliability engineering. He is the co-holder of chair on Risk and Resilience of Complex Systems.

APPENDIX

The research of Zhiguo Zeng is supported by ANR-22-CE10-0004 and the Chair on Risk and Resilience of Complex Systems (EDF, Orange, and SNCF). Zhenling Chen participated in this project as an internship student at LGI.