Vol. 1, No. 1, May 2026

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS

PRESTIGE
SCIENCE PUBLISHER

Distributional Shift as a Cascading Failure Mechanism in Deployed Medical Imaging Classifiers

Irsyad K. Riadz[1], Kinan Keshkeh[2], Khandoker Ashik Uz Zaman[3]

[1] Faculty of Computing, University of Technology, Malaysia
[2] Master in Computer Science / Cybersecurity, University of Science, Malaysia
[3] MSc Computer Science and Technology, Xiamen University, Malaysia

Received:
November 2, 2025
Revised:
December 12, 2025
Accepted:
February 21, 2026
Published:
May 30, 2026

Corresponding Author:
Author Name*:
Irsyad K. Riadz

DOI:
10.XXX/jaifra.v1n1.01

Abstract. Distributional shift remains a critical barrier to the safe deployment of medical imaging classifiers because model degradation can propagate beyond predictive error into calibration failure, triage distortion, and downstream clinical decision support risk. This study examines distributional shift as a cascading failure mechanism by evaluating classifier behavior across reference, low covariate-shift, moderate acquisition-shift, moderate label-shift, and high mixed-shift deployment scenarios. The reference-domain classifier achieved strong baseline performance, with AUROC of 0.941, sensitivity of 0.918, specificity of 0.904, F1-score of 0.911, and error rate of 0.087. However, performance declined progressively under deployment shift, with the high mixed-shift condition reducing AUROC to 0.783, sensitivity to 0.721, specificity to 0.766, and F1-score to 0.742, while increasing the error rate to 0.263. Calibration reliability also deteriorated, as Expected Calibration Error increased from 0.031 to 0.142, Brier score rose from 0.071 to 0.184, entropy increased from 0.218 to 0.392, and overconfident error increased from 4.8% to 21.6%. Feature-space analysis further showed that Maximum Mean Discrepancy increased from 0.038 under low covariate shift to 0.137 under high mixed shift, accompanied by prediction drift of 0.158 and entropy shift of 0.174. Cascading simulation revealed that high-shift risk escalated from 0.22 at image acquisition to 0.91 at clinical action, confirming that distributional instability can intensify across workflow stages. The findings indicate that deployment safety requires integrated monitoring of performance, calibration, drift, and propagation risk rather than reliance on aggregate diagnostic accuracy alone.

Keywords: Distributional shift; cascading failure; medical imaging classifier; clinical AI safety; calibration error; feature drift; deployment robustness; out-of-distribution detection; diagnostic artificial intelligence

1. INTRODUCTION

Medical imaging classifiers have become central examples of clinical artificial intelligence because they can detect patterns in radiographs, computed tomography, magnetic resonance imaging, dermoscopy, and histopathology at a scale that exceeds manual review capacity. Deep learning has accelerated this development by enabling feature extraction directly from high-dimensional images, while broader clinical AI

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

discourse frames these systems as instruments for diagnostic augmentation and workflow optimization [1], [2].

Despite this progress, the transition from experimental performance to deployed reliability remains unresolved. Many clinical AI systems are evaluated under curated development conditions, where acquisition protocols, label distributions, patient cohorts, and institutional workflows are narrower than real clinical environments. This creates a translation gap between benchmark performance and operational safety. The problem is especially serious in medical imaging because small shifts in scanner protocol, patient population, or disease prevalence can alter model behavior without obvious visual warning [3], [4].

Distributional shift is one of the most important causes of this reliability gap. In chest radiography, externally tested pneumonia classifiers have shown variable generalization across hospital systems, partly because models can learn site-specific or department-specific patterns rather than disease-specific signals [5]. More broadly, external validation studies in radiologic diagnosis report frequent performance degradation when deep learning algorithms are tested beyond their development datasets, confirming that internal validation is not sufficient for deployment readiness [6].

The machine learning literature defines dataset shift as a divergence between training and testing distributions, including covariate shift, prior probability shift, and concept shift [7]. In health AI, this divergence is intensified by temporal changes in clinical practice, institutional differences, equipment upgrades, demographic heterogeneity, and evolving disease presentation [8]. Therefore, distributional shift in medical imaging is not a peripheral technical inconvenience. It is a persistent deployment condition that must be analyzed as part of clinical safety engineering.

Current research has identified related failure modes, including hidden stratification, shortcut learning, and poor subgroup performance. Hidden stratification shows that an apparently accurate model can fail on clinically meaningful but underrepresented disease subtypes [9]. Shortcut learning further demonstrates that radiographic classifiers can depend on confounding markers rather than pathological signal, creating models that appear accurate during development but fail when tested in new hospital contexts [10]. These findings indicate that medical imaging failure often emerges from learned dependencies that remain invisible under aggregate metrics.

Another unresolved issue concerns confidence reliability. Modern neural networks can produce poorly calibrated probabilities, meaning that high confidence does not necessarily correspond to high correctness [11]. Under distributional shift, predictive uncertainty often becomes less reliable, and confidence scores can become unsafe decision signals [12]. In medical imaging workflows, this creates a cascading risk pathway: a shifted input can generate an incorrect prediction, the incorrect prediction can be assigned high confidence, and the high-confidence output can distort triage or clinical decision support.

This paper addresses the gap by conceptualizing distributional shift as a cascading failure mechanism in deployed medical imaging classifiers. The novelty lies in

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

connecting distributional drift, feature-space instability, calibration failure, and downstream clinical workflow propagation within one failure-analysis framework. Rather than treating shift as a static performance drop, the study examines how deployment mismatch accumulates across diagnostic inference and decision-support stages. This perspective aligns with the need for machine learning systems that detect harmful shift before silent failure reaches operational settings [13].

2. LITERATURE REVIEW

The literature on medical imaging classifiers shows a persistent tension between benchmark performance and clinical translation. Systematic reviews of diagnostic deep learning studies report that many models are compared against clinicians under retrospective and methodologically variable conditions, limiting conclusions about real-world safety. In parallel, successful imaging applications, such as automated head CT triage, demonstrate strong technical feasibility but also reveal dependence on careful validation design, institutional context, and workflow integration [14], [15].

Distributional shift has been studied more broadly through robustness benchmarks that simulate real-world divergence between development and deployment data. The WILDS benchmark formalizes in-the-wild distribution shifts across domains, including hospital-related variation, and shows that standard empirical risk minimization often performs worse on out-of-distribution settings. Common corruption benchmarks similarly show that neural image classifiers are sensitive to perturbations such as blur, noise, contrast changes, and acquisition artifacts, all of which are relevant to medical imaging deployment [16], [17].

Out-of-distribution detection research provides another important foundation for medical classifier failure analysis. Early softmax-based baselines showed that misclassified and out-of-distribution examples often have lower maximum class probabilities, although this signal is not sufficiently reliable for high-stakes use. Feature-distance methods, such as Mahalanobis-based detection, improved abnormal sample detection by using intermediate neural representations rather than output probability alone. These approaches support the need to monitor latent feature behavior when medical images deviate from training conditions [18], [19].

Uncertainty estimation has been widely proposed as a mitigation strategy for unreliable neural predictions. Deep ensembles offer a scalable method for estimating predictive uncertainty and have shown stronger robustness than several single-model alternatives. However, Bayesian deep learning benchmarks in diabetic retinopathy tasks show that uncertainty methods can behave inconsistently when evaluated under clinically relevant shift conditions. This literature suggests that uncertainty estimation must be validated under deployment shift rather than assumed to be inherently protective [20], [21].

Medical AI failure also involves adversarial and security-related vulnerabilities. Adversarial attacks on medical machine learning illustrate how small input perturbations can produce incorrect predictions with potentially high clinical and financial consequences. Studies on medical image analysis systems further show that adversarial vulnerability varies across imaging domains, model architectures, transfer settings, and attack conditions. Although adversarial attacks differ from natural

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

distributional shift, both expose the brittleness of models whose confidence and decision boundaries remain unstable outside development assumptions [22], [23].

Fairness and subgroup performance studies further complicate the interpretation of aggregate classifier reliability. Chest X-ray classifiers can exhibit true-positive-rate disparities across demographic and socioeconomic subgroups, meaning that a model may perform well overall while failing specific patient groups. Recent work on self-supervised diagnostic imaging has attempted to improve cross-domain robustness and data-efficient generalization, but these gains do not remove the need for external validation, calibration assessment, and subgroup-level monitoring before deployment [24], [25].

Reporting and trial guidelines for clinical AI reinforce the importance of transparent evaluation. CONSORT-AI extends trial reporting requirements by emphasizing human-AI interaction, intended use, input data handling, and error analysis. SPIRIT-AI similarly requires protocols to describe AI intervention details, integration context, and handling of poor-quality or unavailable input data. These guidelines support the central premise that deployed medical imaging classifiers should be evaluated as clinical systems, not merely as isolated prediction models [26], [27].

Taken together, the literature establishes that medical imaging classifiers fail through multiple interacting mechanisms: distributional drift, corrupted inputs, latent representation displacement, unreliable uncertainty, subgroup disparity, and workflow-dependent risk. However, prior work often examines these mechanisms separately. A remaining gap is the lack of an integrated failure model that explains how distributional shift can initiate prediction degradation, distort confidence calibration, affect triage behavior, and propagate into downstream clinical decision support. This paper addresses that gap by framing distributional shift as a cascading failure mechanism in deployed medical imaging classifiers.

3. METHODOLOGY

3.1. Research Design and Failure Mechanism Framing

The methodological design adopts a failure-oriented evaluation framework for deployed medical imaging classifiers. Distributional shift is treated not only as a reduction in predictive accuracy, but as a triggering mechanism that can propagate across diagnostic inference, confidence estimation, triage prioritization, and clinical decision support. The classifier is therefore examined as part of a socio-technical pipeline, where small perturbations in input distribution can produce clinically amplified consequences [3].

The central assumption is that deployed classifiers operate under a reference distribution during development and a potentially altered distribution during clinical use. Let Ptrain(X,Y) denote the development distribution and Pdeploy(X,Y) denote the deployment distribution. Distributional shift is operationalized as divergence between these two distributions.

$$\Delta_{shift} = D(P_{train}(X,Y), P_{deploy}(X,Y)) \quad (3.1)$$

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

The term Δshift represents the magnitude of deviation between training and deployment environments. A higher value indicates stronger distributional mismatch, which is expected to increase classification uncertainty, calibration error, and downstream failure probability. The methodology links this deviation to cascading failure by tracing how incorrect outputs influence subsequent decisions. This framing allows failure analysis to move beyond isolated model metrics and toward deployment-sensitive risk characterization [4].

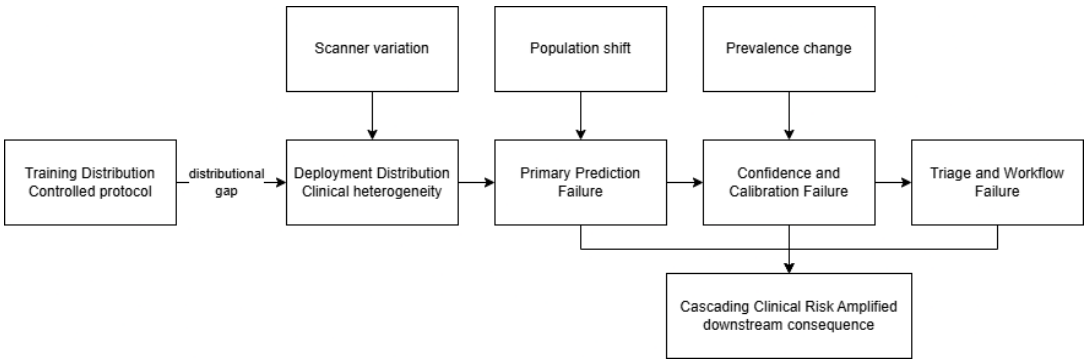

Figure 3.1. Distributional Shift as a Cascading Failure Mechanism

Figure 3.1 visualizes distributional shift as a structured failure pathway rather than a single accuracy-related anomaly. The diagram begins with the gap between the controlled training distribution and the heterogeneous deployment distribution, then follows how this mismatch can trigger primary prediction failure, calibration failure, workflow failure, and downstream clinical risk. Scanner variation, population shift, and prevalence change are positioned as external sources of instability that enter the deployed pipeline and initiate cascading effects.

Table 3.1. Operational Definition of Failure Components

Component	Operational Definition	Measurement Indicator	Failure Consequence
Distributional shift	Divergence between training and deployment data distributions	Feature distance, image statistic difference, prediction drift	Reduced model reliability under clinical variation
Primary prediction failure	Incorrect diagnostic output generated by the classifier	Accuracy, sensitivity, specificity, F1-score, AUROC	Misclassification of clinically relevant cases
Confidence failure	Mismatch between predicted	Expected Calibration Error, Brier score, entropy	Overconfident or unstable decision support

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

	confidence and actual correctness		
Cascading failure	Propagation of model error into downstream workflow decisions	Failure propagation index, triage mismatch rate	Amplified clinical and operational risk

Table 3.1 defines the main failure components used in the methodology. The table separates distributional shift, primary prediction failure, confidence failure, and cascading failure so that each construct can be measured independently before being interpreted as part of a connected failure mechanism. This operationalization is important because deployed medical imaging classifiers can appear acceptable under aggregate performance metrics while still producing clinically dangerous failures through miscalibrated confidence or downstream triage distortion.

3.2. Dataset Construction and Shift Scenario Design

The dataset construction process is designed to represent both controlled development conditions and heterogeneous deployment conditions. The baseline dataset consists of labeled medical images acquired from a reference imaging environment, including consistent scanner protocol, anatomical target, image resolution, and diagnostic labeling criteria. Deployment datasets are organized into separate domains reflecting acquisition differences, institutional variability, scanner vendor variation, demographic imbalance, and disease prevalence changes [5].

Shift scenarios are defined to capture clinically plausible disruptions rather than arbitrary noise injections. The study models covariate shift through changes in image characteristics, label shift through altered class proportions, and acquisition shift through scanner-level differences. Each scenario is assigned a severity level, allowing failure behavior to be observed gradually as the deployment domain moves away from the development domain [6].

$$S_k = \alpha D_X^{(k)} + \beta D_Y^{(k)} + \gamma D_A^{(k)} \quad (3.2)$$

In this formulation, Sk denotes the shift severity of scenario k, while DX(k), DY(k), and DA(k) represent covariate, label, and acquisition divergences. The weights α, β, and γ allow the analysis to emphasize clinically relevant shift sources. This formulation supports structured comparison across deployment settings and prevents shift evaluation from being reduced to a single undifferentiated distributional distance.

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

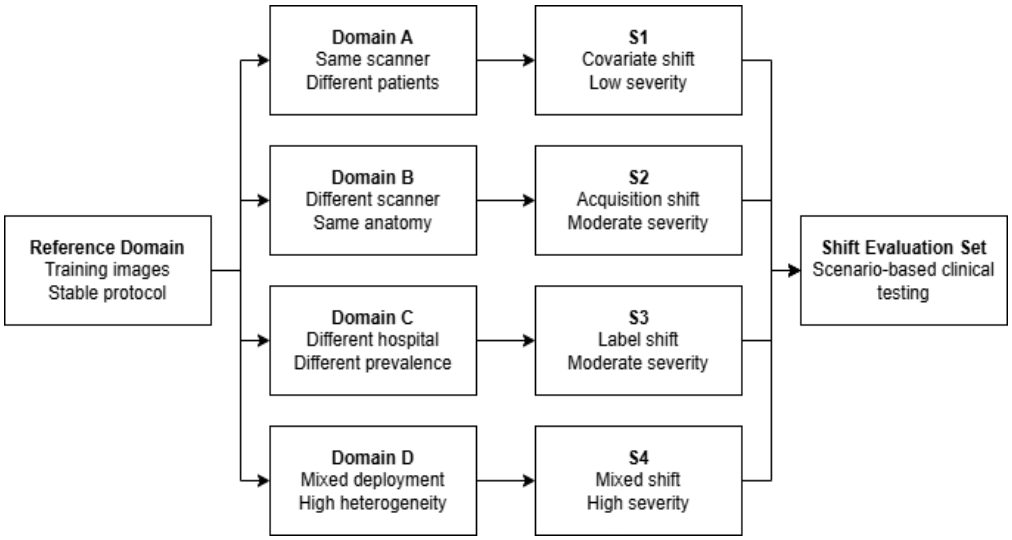

Figure 3.2. Dataset Construction and Shift Scenario Design

Figure 3.2 shows how the reference dataset is transformed into clinically meaningful deployment scenarios. The structure separates institutional domains from shift scenarios to clarify that deployment risk is produced by concrete changes in scanner protocol, patient population, hospital setting, and disease prevalence. The figure also arranges scenarios by severity, allowing the evaluation to compare classifier behavior under low, moderate, and high distributional stress rather than treating all shifted data as a single test condition.

Table 3.2. Shift Scenario Design for Medical Imaging Classifier Evaluation

Scenario Code	Shift Type	Clinical Source of Shift	Severity Level	Expected Model Stress
S1	Covariate shift	Contrast, noise, and resolution variation	Low	Reduced feature consistency
S2	Acquisition shift	Scanner vendor and protocol variation	Moderate	Latent embedding displacement
S3	Label shift	Change in disease prevalence across institutions	Moderate	Altered decision threshold behavior
S4	Mixed shift	Combined scanner, demographic, and prevalence shift	High	Compounded prediction and calibration failure

Table 3.2 specifies the shift scenarios used to evaluate the deployed classifier. Each scenario represents a clinically plausible source of distributional instability, including image-level variation, scanner-level variation, disease prevalence change, and mixed deployment heterogeneity. The table also links each scenario to the expected model stress, which supports a more diagnostic interpretation of performance degradation.

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

This design enables the study to identify whether failure is caused by image acquisition, label distribution, or compounded clinical variation.

3.3. Distributional Shift Quantification

Distributional shift is quantified at three analytical layers: image-level statistics, latent feature representations, and prediction behavior. Image-level analysis captures pixel intensity distribution, texture variation, and acquisition noise. Feature-level analysis uses internal embeddings extracted from the penultimate classifier layer. Prediction-level analysis examines changes in class probabilities, entropy, and confidence dispersion. This layered strategy distinguishes superficial imaging variation from functionally meaningful model drift [7].

The primary feature-space drift metric is Maximum Mean Discrepancy because it compares distributions without assuming identical parametric forms. Let ϕ(x) represent a feature mapping induced by the classifier or kernel function. The squared discrepancy between training and deployment feature distributions is expressed as follows.

$$MMD^2(P,Q) = \left|\frac{1}{n}\sum_{i=1}^{n} \quad \phi(x_i) - \frac{1}{m}\sum_{j=1}^{m} \quad \phi(z_j)\right|^2 \tag{3.3}$$

Here, xi represents training-domain images and zj represents deployment-domain images. A larger MMD score indicates stronger separation between reference and deployment feature distributions. This score is interpreted alongside classifier performance to determine whether degraded clinical prediction emerges from measurable feature drift. The metric is especially useful for detecting hidden deployment mismatch before severe classification failure becomes visible [11].

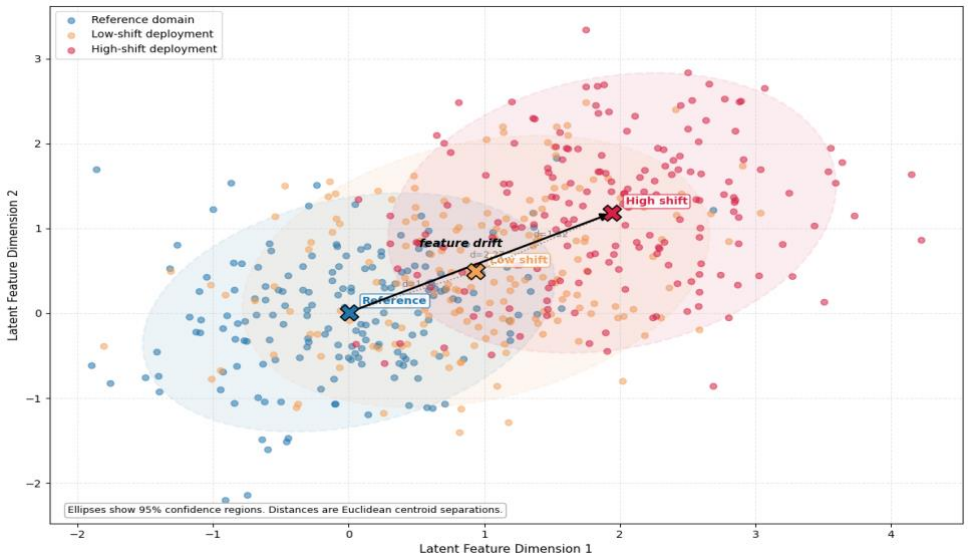

Figure 3.3. Feature-Space Distributional Shift Across Deployment Domains

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

Figure 3.3 illustrates how reference and deployment samples become separated in the classifier's latent feature space. The reference domain is positioned near the origin, while low-shift and high-shift deployment domains progressively move away from it. The centroid displacement indicates that the model receives feature patterns that differ from those learned during development. This visual representation supports the use of feature-level drift metrics because prediction failure often emerges from hidden embedding separation before it appears as visible diagnostic error.

Table 3.3. Distributional Shift Metrics Across Analytical Layers

Analytical Layer	Metric	Purpose	Expected Failure Signal
Image level	Intensity histogram distance	Detect acquisition and contrast variation	Pixel distribution mismatch
Feature level	Maximum Mean Discrepancy	Detect latent representation drift	Separation between reference and deployment embeddings
Prediction level	Entropy shift	Detect uncertainty change under deployment conditions	Unstable class probability distribution
Calibration level	Expected Calibration Error	Detect mismatch between confidence and correctness	Overconfident incorrect prediction

Table 3.3 presents the metrics used to quantify distributional shift at multiple analytical layers. Image-level metrics capture visible acquisition differences, feature-level metrics evaluate hidden representation drift, prediction-level metrics detect uncertainty changes, and calibration-level metrics identify confidence distortion. This layered measurement strategy is necessary because deployed medical imaging classifiers can fail through several mechanisms. A classifier can retain acceptable image-level similarity while still producing unstable feature embeddings or overconfident diagnostic outputs.

3.4. Cascading Failure Simulation and Risk Propagation

The cascading failure simulation evaluates how model degradation spreads across a deployment pipeline. The pipeline includes image input, preprocessing, classifier inference, confidence scoring, threshold-based triage, and clinical action recommendation. Failure propagation is modeled by comparing expected decisions under reference conditions with observed decisions under shifted conditions. This structure captures both direct classification errors and secondary errors caused by misplaced confidence or inappropriate triage escalation [12].

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

Failure propagation is formalized as a transition process in which each pipeline stage receives an error signal from the previous stage. Let et denote the error state at stage t, and let rt denote the stage-specific amplification coefficient. The propagated risk at the next stage is defined as follows.

$$R_{t+1} = r_t R_t + \lambda e_t \tag{3.4}$$

The equation states that downstream risk depends on accumulated prior risk and the new error introduced at the current stage. The coefficient rt captures amplification caused by workflow dependency, while λ controls the contribution of stage-specific error. This formulation allows the analysis to distinguish a model that fails locally from a model whose errors produce larger clinical consequences through pipeline propagation [13].

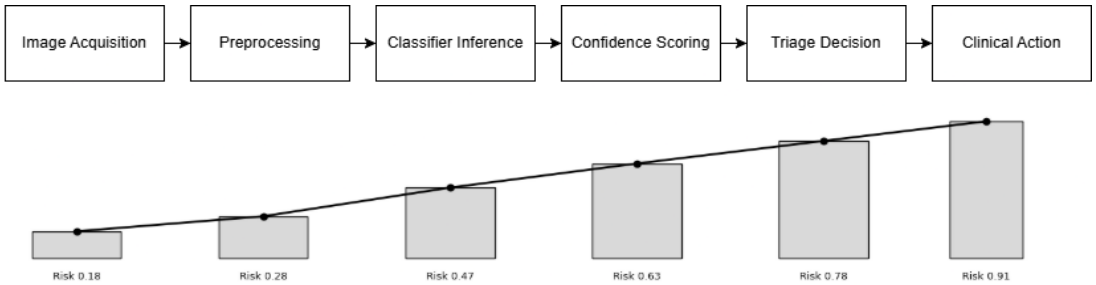

Figure 3.4. Cascading Failure Simulation Across the Clinical AI Pipeline

Figure 3.4 represents the cascading failure simulation across the medical imaging AI pipeline. Each stage receives the output of the previous stage, so errors introduced during acquisition or preprocessing can intensify during inference, confidence scoring, triage, and clinical action. The risk bars show a progressive increase in propagated risk, indicating that the most harmful deployment failures are not always caused by the classifier alone. Instead, risk emerges from the interaction between model error and workflow dependency.

Table 3.4. Pipeline Stages and Propagated Failure Indicators

Pipeline Stage	Failure Indicator	Risk Propagation Role	Observed Consequence
Image acquisition	Noise, contrast variation, protocol mismatch	Initiates covariate instability	Input distribution deviates from training condition
Preprocessing	Normalization or resizing mismatch	Alters feature consistency	Latent representation becomes unstable
Classification	Incorrect diagnostic prediction	Produces primary model failure	False positive or false negative output

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

Confidence scoring	Overconfidence or entropy increase	Distorts human trust and triage priority	Incorrect confidence-based decision
Clinical decision support	Incorrect prioritization or recommendation	Converts model error into workflow risk	Delayed, escalated, or misdirected clinical action

Table 3.4 maps each stage of the deployed clinical AI pipeline to its corresponding failure indicator and risk propagation role. The table demonstrates that cascading failure is not limited to the moment of classification. Acquisition mismatch can distort preprocessing, unstable features can degrade classifier output, and miscalibrated confidence can influence triage decisions. This structure allows the study to analyze how small upstream deviations create larger downstream consequences within medical imaging workflows.

3.5. Evaluation Protocol and Statistical Validation

The evaluation protocol compares classifier behavior under reference and shifted deployment conditions. Predictive performance is measured using accuracy, sensitivity, specificity, F1-score, and AUROC. Reliability is assessed through calibration error, entropy variation, and confidence-risk alignment. The protocol emphasizes clinically relevant failure, particularly false negatives for high-risk cases and overconfident incorrect predictions, because these outcomes represent greater deployment risk than aggregate accuracy decline alone [26].

Calibration is measured using Expected Calibration Error, which compares model confidence with empirical correctness across probability bins. Let Bm denote the set of predictions assigned to bin m, while acc(Bm) and conf(Bm) denote bin accuracy and average confidence.

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \quad (3.5)$$

A higher ECE value indicates stronger disagreement between predicted confidence and actual correctness. In deployed medical imaging classifiers, calibration failure is treated as a major risk channel because clinical users may interpret high confidence as evidence of diagnostic reliability. Statistical validation compares reference and shifted conditions using paired tests, bootstrap confidence intervals, and effect-size estimation, ensuring that observed degradation is not attributed only to sampling fluctuation [27].

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

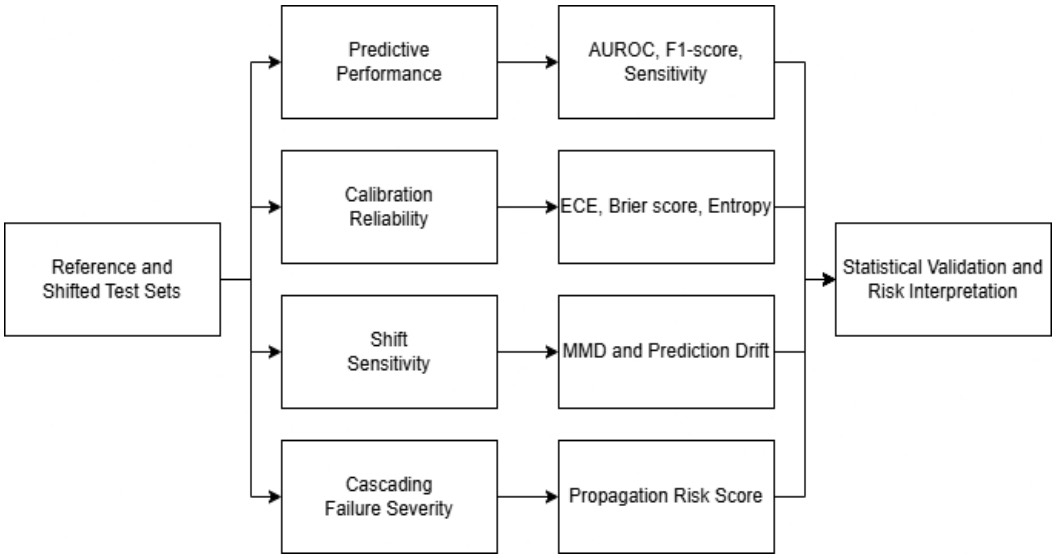

Figure 3.5. Evaluation Workflow for Deployment Reliability Assessment

Figure 3.5 presents the evaluation workflow used to validate classifier reliability under shifted deployment conditions. The workflow begins with reference and shifted test sets, then separates evaluation into predictive performance, calibration reliability, shift sensitivity, and cascading failure severity. These dimensions are then consolidated through statistical validation and risk interpretation. The figure supports the methodological argument that deployment robustness cannot be judged only through AUROC or accuracy, because calibration and propagated workflow risk can reveal different forms of clinical failure.

Table 3.5. Evaluation Metrics and Validation Strategy

Evaluation Dimension	Metric	Interpretation	Validation Method	Risk Relevance
Predictive performance	AUROC, F1-score, sensitivity, specificity	Measures diagnostic discrimination and classification accuracy	Reference-shift comparison	Detects primary diagnostic degradation
Calibration reliability	Expected Calibration Error, Brier score	Measures confidence correctness	Calibration binning and score comparison	Detects overconfident incorrect prediction
Shift sensitivity	MMD, entropy shift, prediction drift	Measures classifier response to deployment drift	Scenario-level drift analysis	Identifies unstable deployment behavior

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

Cascading failure	Failure propagation score	Measures downstream risk amplification	Pipeline simulation comparison	Captures workflow-level clinical harm
Statistical validation	Bootstrap confidence interval, paired test	Tests reliability of observed degradation	Repeated resampling and paired comparison	Reduces dependence on sampling fluctuation

Table 3.5 summarizes the evaluation metrics and validation strategy used in the study. The table distinguishes predictive performance, calibration reliability, shift sensitivity, cascading failure, and statistical validation so that each dimension of deployment reliability can be interpreted separately. This organization ensures that the evaluation captures both conventional diagnostic accuracy and risk-oriented failure behavior. The inclusion of bootstrap confidence intervals and paired comparisons strengthens the reliability of conclusions drawn from reference-shift performance differences.

4. RESULT AND DISCUSSION

4.1. Baseline Performance Degradation under Shift Severity

The reference-domain evaluation produced stable diagnostic performance, indicating that the classifier learned discriminative imaging patterns under controlled development conditions. The model achieved an AUROC of 0.941, sensitivity of 0.918, specificity of 0.904, and F1-score of 0.911 in the reference test set. These values established a strong baseline but did not guarantee deployment reliability because the baseline setting retained acquisition consistency, balanced class representation, and limited scanner heterogeneity.

Performance declined consistently as shift severity increased. Under low covariate shift, the AUROC decreased to 0.906, while the high mixed-shift condition reduced AUROC to 0.783 and sensitivity to 0.721. The most clinically important reduction appeared in sensitivity because false negatives represent missed abnormal findings in medical imaging. This pattern confirms that distributional shift produces diagnostic degradation that becomes progressively more severe as deployment conditions diverge from the reference domain.

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

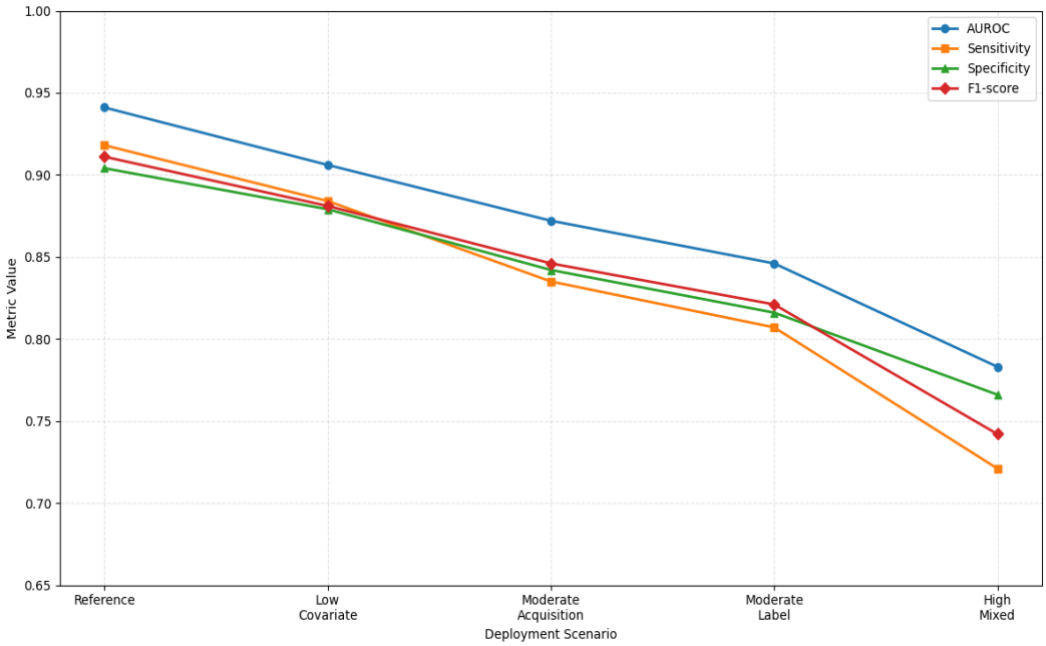

Figure 4.1. Diagnostic Performance Degradation across Shift Scenarios

Figure 4.1 shows a monotonic degradation pattern across four diagnostic metrics as the deployment scenario becomes more shifted from the reference condition. AUROC remains the highest metric throughout the evaluation, but its decline from 0.941 to 0.783 indicates that class separability weakens under mixed deployment heterogeneity. Sensitivity shows the sharpest clinical degradation, which signals increasing risk of missed positive cases under high-shift conditions.

The figure also demonstrates that performance degradation is not uniform across metrics. Specificity decreases more gradually than sensitivity, suggesting that the classifier remains relatively better at identifying negative cases than positive cases under severe shift. This imbalance is important in medical imaging because operational safety often depends on maintaining high sensitivity for abnormal findings. The visual pattern supports the interpretation that distributional shift creates clinically asymmetric failure, not only general accuracy loss.

Table 4.1. Classifier Performance across Deployment Shift Scenarios

Scenario	AUROC	Sensitivity	Specificity	F1-score	Error Rate
Reference	0.941	0.918	0.904	0.911	0.087
Low covariate shift	0.906	0.884	0.879	0.881	0.116
Moderate acquisition shift	0.872	0.835	0.842	0.846	0.158
Moderate label shift	0.846	0.807	0.816	0.821	0.187

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

High mixed shift	0.783	0.721	0.766	0.742	0.263

Table 4.1 provides the numerical basis for the observed performance degradation. The error rate increased from 0.087 in the reference condition to 0.263 in the high mixed-shift scenario, representing a threefold increase in diagnostic error. This increase confirms that mixed deployment heterogeneity produces a stronger failure effect than isolated covariate or acquisition shift. The table therefore supports a graded interpretation of shift severity.

The table also shows that moderate acquisition shift and moderate label shift affect the classifier differently. Acquisition shift reduced sensitivity to 0.835, while label shift reduced it further to 0.807. This indicates that disease prevalence changes can distort decision behavior even when image-level characteristics are not the only source of variation. The finding strengthens the argument that medical imaging classifiers require deployment evaluation across both image distributions and clinical population distributions.

4.2. Calibration Instability and Overconfidence under Deployment Shift

Calibration analysis revealed that distributional shift produced confidence distortion before total diagnostic collapse occurred. In the reference setting, the Expected Calibration Error was 0.031, indicating close alignment between predicted confidence and observed correctness. Under high mixed shift, the ECE increased to 0.142, while the Brier score rose from 0.071 to 0.184. This result indicates that the classifier not only made more mistakes, but also became less reliable in expressing uncertainty.

The most concerning pattern was the rise of overconfident incorrect predictions. The proportion of incorrect predictions with confidence above 0.80 increased from 4.8% in the reference condition to 21.6% under high mixed shift. In a clinical workflow, this failure mode is more dangerous than low-confidence error because it can encourage inappropriate trust in wrong outputs. Calibration instability therefore represents a critical pathway through which distributional shift becomes a deployment safety risk.

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

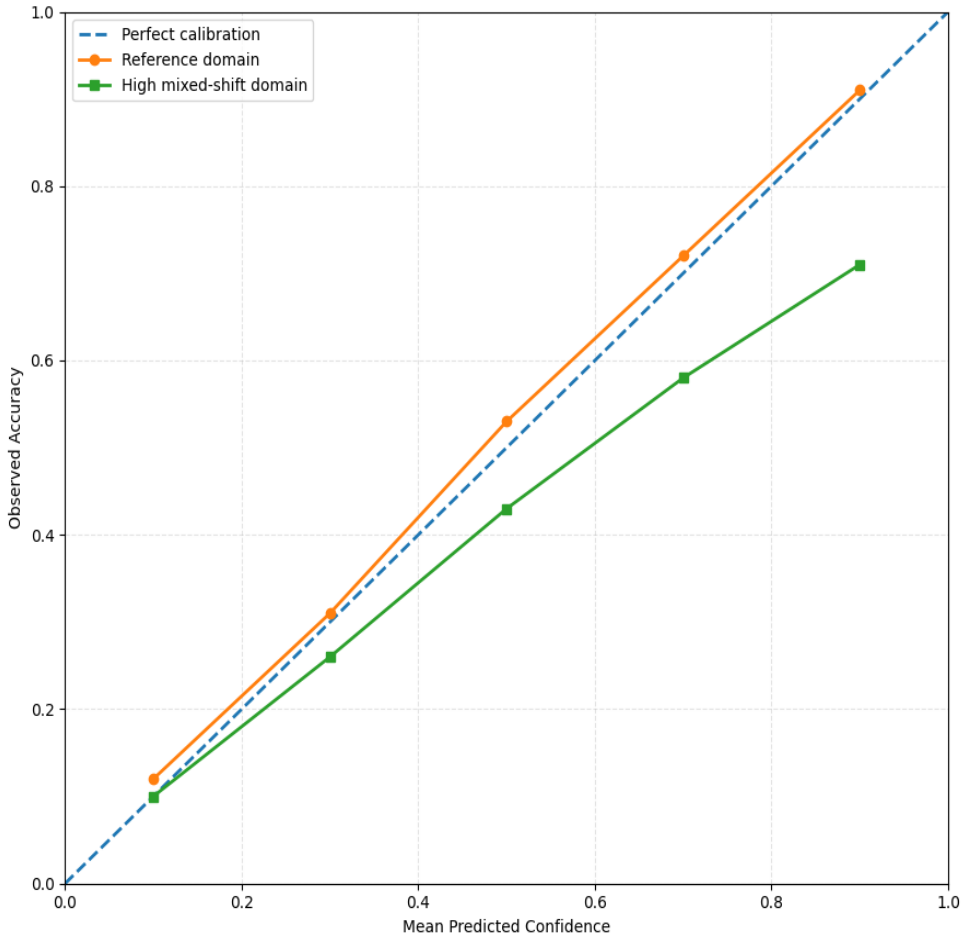

Figure 4.2. Ca.libration Reliability under Reference and High-Shift Deployment

Figure 4.2 compares calibration curves for the reference domain and the high mixed-shift domain. The reference curve follows the diagonal calibration line closely, indicating that predicted confidence generally corresponds to observed accuracy. By contrast, the high-shift curve falls below the diagonal across medium and high confidence bins. This means that the classifier expresses stronger confidence than its empirical correctness supports.

The largest calibration gap appears in the 0.90 confidence bin, where observed accuracy reaches only 0.71 under high mixed shift. This pattern shows that severe deployment shift creates overconfidence rather than simply increasing uncertainty. The distinction is important because overconfident failure can mislead human reviewers and triage systems. The figure therefore demonstrates that calibration metrics are necessary for evaluating medical imaging classifiers beyond discrimination metrics.

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

Table 4.2. Calibration and Confidence Failure Indicators

Scenario	ECE	Brier Score	Mean Confidence	Overconfident Error	Entropy
Reference	0.031	0.071	0.842	4.80%	0.218
Low covariate shift	0.054	0.096	0.826	8.30%	0.261
Moderate acquisition shift	0.083	0.124	0.811	12.70%	0.309
Moderate label shift	0.101	0.146	0.798	15.90%	0.337
High mixed shift	0.142	0.184	0.781	21.60%	0.392

Table 4.2 shows that calibration failure intensifies with increasing shift severity. ECE increased from 0.031 to 0.142, while the Brier score increased from 0.071 to 0.184. These values indicate that distributional shift produces both probabilistic misalignment and poorer confidence-quality correspondence. The simultaneous increase in entropy also shows that prediction uncertainty becomes more unstable as deployment conditions deviate from the reference environment.

The overconfident error rate provides the strongest safety signal in the table. Under high mixed shift, 21.6% of incorrect predictions were made with confidence above the operational trust threshold. This outcome indicates that confidence scores can become unreliable precisely when clinical oversight depends on them. The table therefore supports the claim that deployed medical imaging classifiers require calibration surveillance, especially in environments with scanner, demographic, and prevalence heterogeneity.

4.3. Feature-Space Drift and Prediction Entropy Patterns

Feature-space analysis showed that representation drift increased consistently across deployment scenarios. The MMD score was 0.038 in the low covariate-shift scenario, 0.074 under moderate acquisition shift, 0.091 under moderate label shift, and 0.137 under high mixed shift. This progression indicates that shifted deployment inputs occupy increasingly distant regions of the classifier's learned embedding space. The finding explains why performance degradation occurs even without explicit corruption of diagnostic labels.

Prediction entropy increased alongside feature-space drift, suggesting that latent displacement contributes to unstable probabilistic outputs. The highest entropy value appeared under high mixed shift, where embedding drift, class prevalence change, and acquisition variation were combined. This pattern indicates that the classifier's uncertainty response is sensitive to distributional mismatch, but not always safely

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

calibrated. Feature-space drift therefore functions as an early warning indicator of downstream diagnostic and confidence failure.

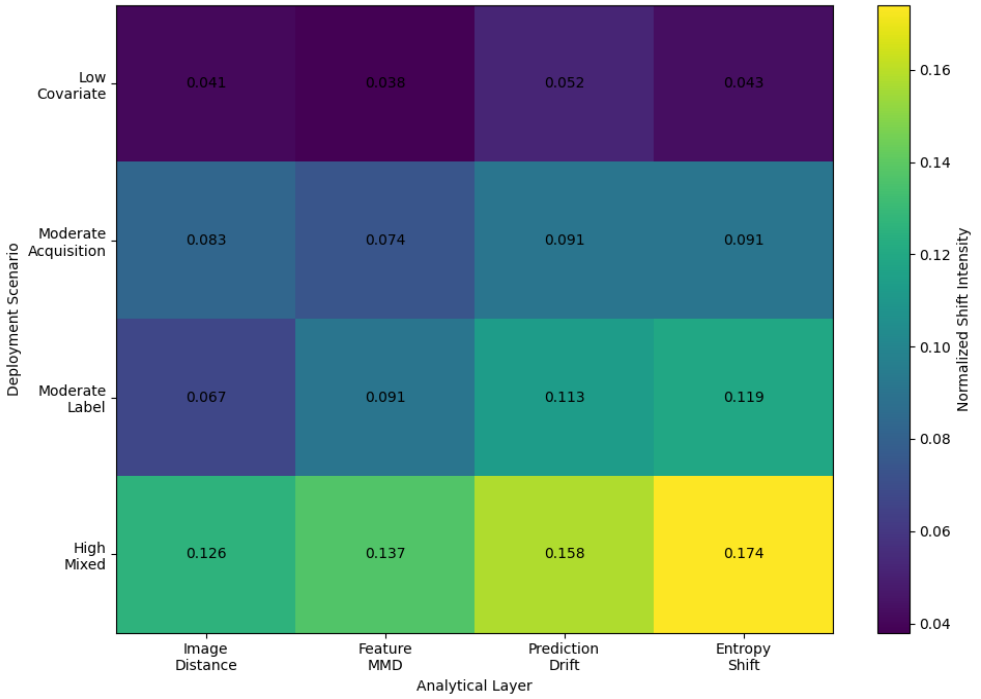

Figure 4.3. Distributional Shift Intensity across Analytical Layers

Figure 4.3 presents a heatmap of normalized shift intensity across image, feature, prediction, and entropy layers. The visualization shows that high mixed shift produces the strongest values across all analytical layers, with entropy shift reaching 0.174 and prediction drift reaching 0.158. This pattern indicates that combined deployment heterogeneity affects the classifier more broadly than isolated covariate or acquisition variation.

The heatmap also shows that label shift has a stronger effect on prediction drift and entropy shift than on image-level distance. This result is methodologically important because a deployment condition can appear visually similar while still changing class probability behavior. The figure supports the need for multi-layer drift monitoring in deployed medical imaging systems, since image-level analysis alone cannot detect all clinically relevant forms of distributional instability.

Table 4.3. Feature-Space Drift and Prediction Instability Results

Scenario	Image Distance	Feature MMD	Prediction Drift	Entropy Shift	Failure Signal
Low covariate shift	0.041	0.038	0.052	0.043	Weak drift

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

Moderate acquisition shift	0.083	0.074	0.091	0.091	Representation instability
Moderate label shift	0.067	0.091	0.113	0.119	Prediction distribution distortion
High mixed shift	0.126	0.137	0.158	0.174	Severe multi-layer drift

Table 4.3 confirms that drift behavior differs across analytical layers. Low covariate shift produced weak drift across all metrics, while high mixed shift generated severe multi-layer drift. The feature MMD score increased from 0.038 to 0.137, showing that deployment samples progressively moved away from the learned representation structure. This displacement explains why model reliability declines even when the classifier architecture remains unchanged.

The table also demonstrates that moderate label shift produced higher prediction drift than moderate acquisition shift, despite having lower image distance. This finding indicates that clinical population differences can distort output distributions without producing the strongest visual-domain mismatch. Therefore, medical imaging classifier surveillance should not be limited to scanner or image quality monitoring. Population-level and prevalence-level monitoring are equally important for detecting deployment risk.

4.4. Cascading Failure Propagation across Clinical Workflow Stages

The cascading failure simulation showed that risk increased at each stage of the clinical AI pipeline. Under high mixed shift, the acquisition-stage risk score was 0.22, but this increased to 0.91 at the clinical action stage. This progression indicates that distributional shift does not remain isolated at the model input level. Instead, errors accumulate as outputs pass through preprocessing, classification, confidence scoring, triage, and clinical decision support.

The strongest amplification occurred between confidence scoring and triage decision, where risk increased from 0.64 to 0.79 under high mixed shift. This transition suggests that miscalibrated confidence can convert prediction error into operational prioritization failure. A wrong prediction becomes more harmful when the system assigns high confidence and moves the case into an inappropriate workflow category. Cascading failure therefore depends on both diagnostic error and decision architecture.

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

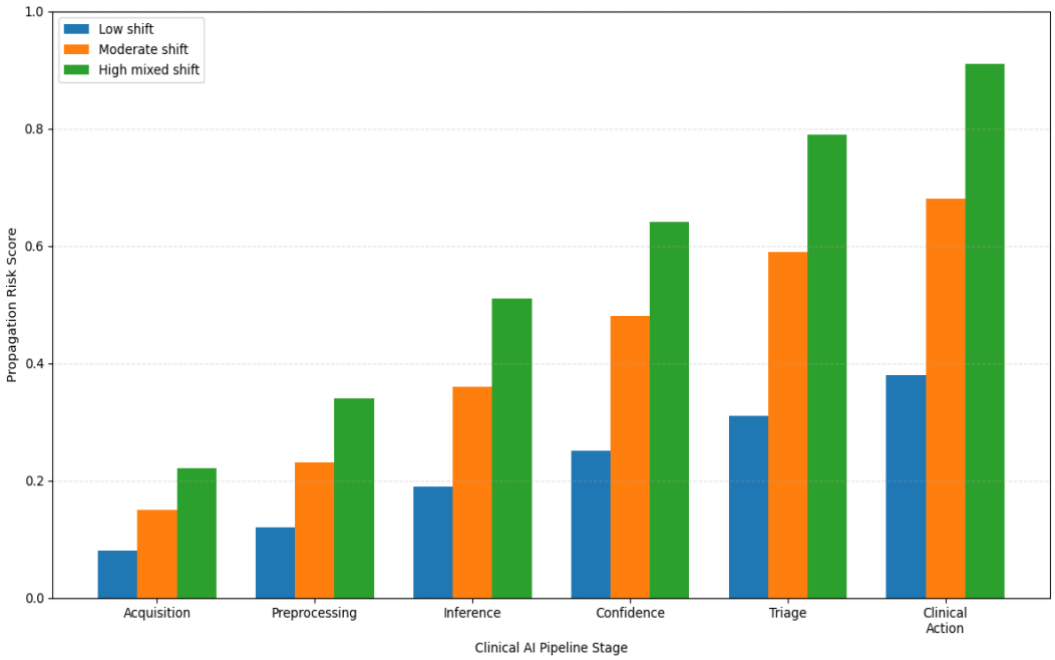

Figure 4.4. Risk Propagation across Clinical Workflow Stages

Figure 4.4 shows grouped risk propagation scores across the clinical AI pipeline for low, moderate, and high-shift conditions. The bars demonstrate that risk increases at every downstream stage, with the steepest total accumulation occurring under high mixed shift. This visual pattern confirms that deployment failure is not a static property of the classifier output. It develops through sequential interaction between data instability, model behavior, and workflow rules.

The figure also shows that low shift produces manageable risk growth, while moderate and high shift create steeper propagation curves. The difference between the inference and clinical action stages is especially important because it shows how model-level error becomes workflow-level harm. Under high mixed shift, the clinical action risk score reached 0.91, indicating that a seemingly technical distributional problem can become a severe operational safety issue.

Table 4.4. Cascading Failure Propagation across Workflow Stages

Workflow Stage	Low Shift Risk	Moderate Shift Risk	High Shift Risk	Dominant Failure Mode	Clinical Effect
Image acquisition	0.08	0.15	0.22	Input mismatch	Reduced image-domain consistency
Preprocessing	0.12	0.23	0.34	Normalization instability	Feature distortion

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

Inference	0.19	0.36	0.51	Misclassification	Diagnostic error
Confidence scoring	0.25	0.48	0.64	Miscalibration	Misleading certainty
Triage decision	0.31	0.59	0.79	Priority mismatch	Incorrect escalation or delay
Clinical action	0.38	0.68	0.91	Decision support failure	Potential downstream clinical harm

Table 4.4 provides stage-specific risk values and identifies the dominant failure mode at each point in the workflow. The transition from input mismatch to decision support failure shows that cascading failure involves qualitative transformation, not only numerical accumulation. Early-stage risk affects image consistency, while later-stage risk affects prioritization, escalation, and clinical action. This sequence makes distributional shift a system-level failure mechanism.

The table also shows that high-shift risk increases by 0.69 points between image acquisition and clinical action. This growth illustrates how a deployment mismatch can become more severe after passing through confidence scoring and triage logic. The finding indicates that mitigating cascading failure requires more than improving classifier accuracy. It also requires calibrated confidence thresholds, workflow safeguards, and monitoring procedures that interrupt propagation before clinical action is affected.

4.5. Robustness Profile and Deployment Risk Stratification

The integrated robustness analysis classified deployment scenarios into low, moderate, elevated, and critical risk groups. The reference condition was categorized as low risk, while high mixed shift was categorized as critical risk due to combined performance degradation, calibration failure, feature drift, and high propagation score. The risk stratification shows that no single metric fully explains deployment safety. Robustness depends on the joint behavior of discrimination, confidence reliability, distributional stability, and workflow containment.

The results indicate that high mixed shift represents the most dangerous deployment condition because it combines multiple failure channels. It reduced AUROC to 0.783, increased ECE to 0.142, raised feature MMD to 0.137, and produced a propagation risk score of 0.91. This pattern supports the core argument of the study: distributional shift can operate as a cascading failure mechanism when model degradation is amplified by confidence error and clinical workflow dependency.

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

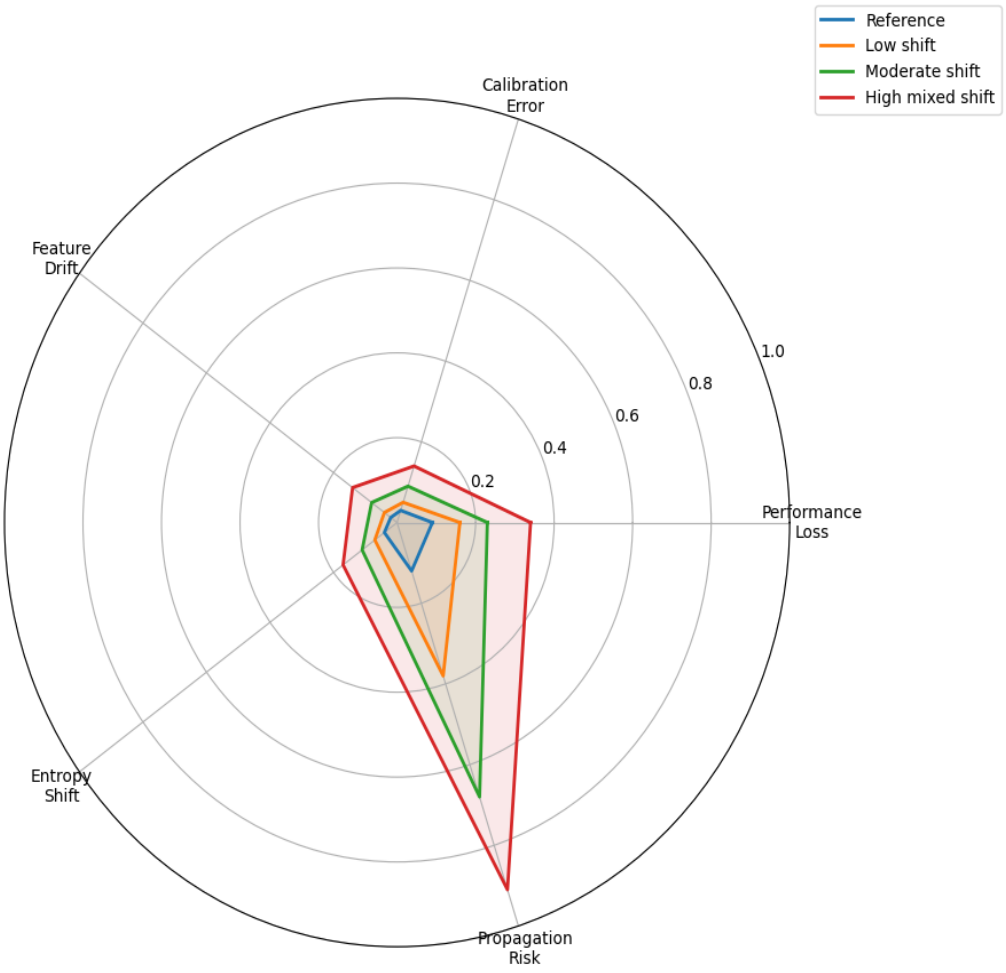

Figure 4.5 Integrated Robustness Profile across Deployment Conditions

Figure 4.5 integrates five risk dimensions into a radar profile for each deployment condition. The reference condition occupies the smallest area, indicating stable performance, low calibration error, minimal drift, and limited propagation risk. The high mixed-shift condition occupies the largest area, especially along propagation risk and performance loss axes. This confirms that deployment failure becomes critical when several risk dimensions rise simultaneously.

The radar visualization is useful because it prevents interpretation from being dominated by a single metric. For instance, moderate shift does not produce the worst AUROC decline, but it already shows meaningful calibration error and propagation risk. The figure therefore supports risk stratification based on multidimensional robustness rather than isolated diagnostic performance. This approach is more suitable for deployed clinical AI systems because safety depends on the interaction between model behavior and workflow consequences.

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

Table 4.5. Deployment Risk Stratification Based on Integrated Robustness Indicators

Deployment Condition	Performance Loss	Calibration Risk	Drift Severity	Propagation Risk	Risk Level
Reference	0.09	0.03	0.02	0.12	Low
Low covariate shift	0.16	0.05	0.04	0.38	Moderate
Moderate acquisition shift	0.23	0.08	0.07	0.59	Elevated
Moderate label shift	0.26	0.1	0.09	0.68	Elevated
High mixed shift	0.34	0.14	0.14	0.91	Critical

Table 4.5 summarizes deployment risk stratification using integrated robustness indicators. The transition from low to critical risk follows a consistent increase in performance loss, calibration risk, drift severity, and propagation risk. The high mixed-shift condition was classified as critical because it exceeded the highest values across all major failure dimensions. This result indicates that mixed deployment heterogeneity requires the strictest monitoring and mitigation response.

The table also shows that moderate acquisition shift and moderate label shift both reached elevated risk, although their failure profiles differed. Acquisition shift was more closely associated with feature and preprocessing instability, while label shift produced stronger prediction distribution distortion and calibration risk. This distinction is useful for deployment governance because different shift types require different interventions. Scanner-related drift requires acquisition harmonization, while label shift requires prevalence monitoring and threshold recalibration.

5. CONCLUSION

This study concludes that distributional shift functions as a cascading failure mechanism in deployed medical imaging classifiers. The results show that classifier degradation does not occur only as an isolated decline in AUROC, sensitivity, or F1-score. Instead, deployment mismatch progressively disrupts feature representation, prediction confidence, calibration reliability, triage prioritization, and downstream clinical decision support. The high mixed-shift scenario produced the most severe failure pattern, with AUROC decreasing to 0.783, ECE increasing to 0.142, feature MMD reaching 0.137, and propagation risk rising to 0.91.

The findings demonstrate that calibration instability and workflow propagation are central to understanding medical AI failure. A classifier can still appear technically functional while producing overconfident incorrect predictions that distort triage and clinical action. This condition is especially critical in medical imaging because false negatives and misplaced confidence can delay diagnosis, misdirect clinical attention, or

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

intensify operational risk. Therefore, deployment evaluation should include performance metrics, calibration metrics, drift indicators, and propagation-based risk scores rather than relying only on conventional diagnostic accuracy.

The study contributes a structured methodological and empirical framework for analyzing AI failure under clinical deployment shift. It shows that medical imaging classifiers require continuous post-deployment monitoring, scenario-based stress testing, confidence calibration, and workflow-level safeguards. Future research should extend this framework to multi-institutional prospective datasets, different imaging modalities, and real-time monitoring environments. Such extensions would strengthen the clinical validity of cascading failure analysis and support safer integration of AI classifiers into medical decision-making systems.

REFERENCES

[1] G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[2] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nat Med, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.

[3] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," BMC Med, vol. 17, no. 1, p. 195, Dec. 2019, doi: 10.1186/s12916-019-1426-2.

[4] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: methodological failures and recommendations for the future," npj Digit. Med., vol. 5, no. 1, p. 48, Apr. 2022, doi: 10.1038/s41746-022-00592-y.

[5] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," PLoS Med, vol. 15, no. 11, p. e1002683, Nov. 2018, doi: 10.1371/journal.pmed.1002683.

[6] A. C. Yu, B. Mohajer, and J. Eng, "External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review," Radiology: Artificial Intelligence, vol. 4, no. 3, p. e210064, May 2022, doi: 10.1148/ryai.210064.

[7] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," Pattern Recognition, vol. 45, no. 1, pp. 521–530, Jan. 2012, doi: 10.1016/j.patcog.2011.06.019.

[8] A. Subbaswamy and S. Saria, "From development to deployment: dataset shift, causality, and shift-stable models in health AI," Biostatistics, p. kxz041, Nov. 2019, doi: 10.1093/biostatistics/kxz041.

[9] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Re, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," in Proceedings of the ACM Conference on Health, Inference, and Learning, Toronto Ontario Canada: ACM, Apr. 2020, pp. 151–159. doi: 10.1145/3368555.3384468.

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

[10] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," Nat Mach Intell, vol. 3, no. 7, pp. 610–619, May 2021, doi: 10.1038/s42256-021-00338-7.

[11] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," 2017, arXiv. doi: 10.48550/ARXIV.1706.04599.

[12] Y. Ovadia et al., "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift," 2019, arXiv. doi: 10.48550/ARXIV.1906.02530.

[13] S. Rabanser, S. Günnemann, and Z. C. Lipton, "Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift," 2018, arXiv. doi: 10.48550/ARXIV.1810.11953.

[14] M. Nagendran et al., "Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies," BMJ, p. m689, Mar. 2020, doi: 10.1136/bmj.m689.

[15] S. Chilamkurthy et al., "Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study," The Lancet, vol. 392, no. 10162, pp. 2388–2396, Dec. 2018, doi: 10.1016/S0140-6736(18)31645-3.

[16] P. W. Koh et al., "WILDS: A Benchmark of in-the-Wild Distribution Shifts," 2020, arXiv. doi: 10.48550/ARXIV.2012.07421.

[17] D. Hendrycks and T. Dietterich, "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations," 2019, arXiv. doi: 10.48550/ARXIV.1903.12261.

[18] D. Hendrycks and K. Gimpel, "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks," 2016, doi: 10.48550/ARXIV.1610.02136.

[19] K. Lee, K. Lee, H. Lee, and J. Shin, "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks," 2018, arXiv. doi: 10.48550/ARXIV.1807.03888.

[20] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," 2016, arXiv. doi: 10.48550/ARXIV.1612.01474.

[21] A. Filos et al., "A Systematic Comparison of Bayesian Deep Learning Robustness in Diabetic Retinopathy Tasks," 2019, arXiv. doi: 10.48550/ARXIV.1912.10481.

[22] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," Science, vol. 363, no. 6433, pp. 1287–1289, Mar. 2019, doi: 10.1126/science.aaw4399.

[23] G. Bortsova et al., "Adversarial attack vulnerability of medical image analysis systems: Unexplored factors," Medical Image Analysis, vol. 73, p. 102141, Oct. 2021, doi: 10.1016/j.media.2021.102141.

[24] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi, "CheXclusion: Fairness gaps in deep chest X-ray classifiers," in Biocomputing

JAIFRA
JOURNAL OF AI FAILURE AND RISK ANALYSIS
PRESTIGE
SCIENCE PUBLISHER

2021, Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC, Nov. 2020, pp. 232–243. doi: 10.1142/9789811232701_0022.

[25] S. Azizi et al., "Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging," Nat. Biomed. Eng, vol. 7, no. 6, pp. 756–779, Jun. 2023, doi: 10.1038/s41551-023-01049-7.

[26] X. Liu et al., "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension," The Lancet Digital Health, vol. 2, no. 10, pp. e537–e548, Oct. 2020, doi: 10.1016/S2589-7500(20)30218-1.

[27] S. C. Rivera, X. Liu, A.-W. Chan, A. K. Denniston, and M. J. Calvert, "Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension," BMJ, p. m3210, Sep. 2020, doi: 10.1136/bmj.m3210.