

Revisiting RIAWELC: Radiograph-Level Auditing Reveals Data Leakage and Class-Selective Domain Shift in Weld Defect Classification

beny beny

beny@unama.ac.id

Universitas Dinamika Bangsa

hendry hendry

Satya Wacana Christian University

ivanna k. timotius

Satya Wacana Christian University

budhi kristianto

Satya Wacana Christian University

Research Article

Keywords:

Posted Date: October 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10519933/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Revisiting RIAWELC: Radiograph-Level Auditing Reveals Data Leakage and Class-Selective Domain Shift in Weld Defect Classification

Beny ^{1,2*}, Hendry ¹, Ivanna K. Timotius³, Budhi Kristianto¹

^{1*}Faculty of Information Technology, Satya Wacana Christian University, Salatiga, Indonesia, 50711.

²Faculty of Computer Science, Dinamika Bangsa University, Jambi, Indonesia, 36138.

³Department of Electronic Engineering, Satya Wacana Christian University, Salatiga, Indonesia, 50711.

*Corresponding author(s). E-mail(s): beny@unama.ac.id;

Abstract

Radiographic weld defect classification with deep learning has advanced through patch-based public datasets, but patch-level evaluation can produce optimistic results when patches from the same source radiograph appear in both training and testing sets. We revisit the RIAWELC dataset. An MD5 audit confirms all 2,443 files in the original test set are byte-identical duplicates of training files, consistent with an independent, concurrent report; previously reported accuracies therefore measure memorization rather than generalization. We construct RIAWELC-RG, a radiograph-level split that removes all duplicates and assigns every patch from a given source radiograph to a single subset. Because leakage correction alone does not reveal domain shift within individual source radiographs, our main contribution is a radiograph-level auditing protocol combining per-radiograph performance analysis, embedding-distance comparison, statistical intensity profiling, and domain-classifier attribution, demonstrated on bam5, an anomalously named radiograph in the test set. Under a fully fine-tuned ConvNeXtV2-T, degradation from bam5 is class-selective: crack accuracy drops to 0.380 while lack of penetration and porosity remain at 1.000. A statistical-feature domain classifier separates bam5 from RRT-domain radiographs almost perfectly (Random Forest area under the receiver operating characteristic curve of 0.9995), with Laplacian sharpness variance the strongest factor, implicating acquisition blur. The pattern reproduces under 5-fold radiograph-level cross-validation and across

thirteen independently trained convolutional (CNN) and Transformer architectures, all showing a positive accuracy gap when bam5 is included (paired t-test, $p < 10^{-6}$). Without radiograph-level evaluation and domain-shift auditing, classifiers trained on datasets like RIAWELC risk mistaking acquisition artifacts for genuine generalization.

Keywords: radiographic weld inspection; data leakage; radiograph-level split; domain shift; class-selective degradation; domain classifier

1 Introduction

Welding inspection is a critical component of industrial quality assurance because defects such as cracks, lack of penetration, and porosity may compromise structural integrity. Radiographic non-destructive testing (NDT) provides an effective way to inspect internal weld conditions, but manual interpretation remains time-consuming and subjective. Automated classification of weld defects using deep learning is therefore an important direction for improving inspection consistency.

Among public datasets in this area, alongside other industrial radiography datasets such as GDXray [1] and WDXI [2], RIAWELC provides 24,407 radiographic image patches resized to 224×224 across four classes (crack, lack of penetration, porosity, no defect), originally distributed into training/validation/testing subsets with a 65/25/10 patch-level split. A key concern with such patch-based datasets is information leakage: because RIAWELC patches are extracted from larger source radiographs, distributing patches from the same radiograph across training and testing subsets can inflate test accuracy through memorization rather than generalization. Sample-level rather than source-level splitting has been shown to inflate reported accuracy across several imaging domains, including optical coherence tomography [3], brain MRI [4], and medical radiography [5], with similar concerns recently recognized in industrial inspection [6], [7].

Beyond leakage, radiograph-level evaluation also exposes domain-shift effects that are hidden under patch-level splits. During preparation we noted that the radiograph ID bam5 was anomalous in its naming relative to the standard RRT-XXR pattern used by all other radiographs, motivating closer scrutiny: constructing otherwise-identical splits with and without bam5 revealed a consistent performance gap across multiple architectures, and visual inspection indicated that bam5 appeared blurrier than typical RRT-series radiographs. This observation was later confirmed quantitatively by sharpness statistics and a domain classifier dominated by Laplacian sharpness variance.

Concurrent to our work, a public GitHub release by Thfmm [8] independently identified duplicate-level leakage in the original RIAWELC and released a deduplicated, radiograph-level split named RIAWELC-RS. We confirm and verify those leakage findings via MD5 hashing on our own re-split. Leakage correction alone, however, does not address a second failure mode in patch-based industrial radiography datasets: acquisition-driven domain shift concentrated in a small number of source

radiographs, invisible to aggregate accuracy and left unexamined by RIAWELC-RS’s deduplication. Our main contribution is a radiograph-level auditing protocol that combines per-radiograph performance analysis, embedding-distance-to-centroid comparison, statistical intensity and texture profiling, and domain-classifier-based feature attribution into a single diagnostic pipeline for surfacing such hidden domain-shift radiographs in patch-based NDT datasets. We demonstrate this protocol through an in-depth case study on RIAWELC: (i) we verify and quantify both duplicate-level and radiograph-level leakage in the original RIAWELC and construct RIAWELC-RG, a leakage-free radiograph-level split; (ii) applying the auditing protocol to RIAWELC-RG, we identify bam5, an anomalously named radiograph, as a class-selective domain-shift case that deduplication alone leaves undetected, with severe degradation concentrated in a single defect class; (iii) the protocol traces the shift to a specific, interpretable mechanism, namely Laplacian sharpness variance, which separates bam5 from the RRT domain with near-perfect accuracy (Random Forest ROC-AUC 0.9995); and (iv) we validate the generality of this finding within RIAWELC through 5-fold radiograph-level cross-validation and a cross-architecture evaluation spanning thirteen CNN and Transformer models. Together, these results show that radiograph-level evaluation must be paired with explicit domain-shift auditing, not leakage correction alone, to avoid mistaking acquisition artifacts for genuine generalization failure.

2 Related Work

2.1 Deep Learning for Radiographic Weld Defect Classification

Deep learning models, particularly CNNs, are widely used for radiographic weld defect classification. The original RIAWELC study [9] reported 93.33% test accuracy with a customized SqueezeNet [10], but high test accuracy alone does not guarantee reliable generalization when the split protocol allows related patches from the same radiograph to appear across subsets. More recent RIAWELC studies have continued to report high accuracy using deeper CNN architectures, including a ResNet50-based model [11] and a VGG-inspired model with visual interpretability [12], without examining whether the original patch-level split introduces leakage [13]. Beyond CNNs, recent work has evaluated Vision Transformers for industrial welding defect detection [14], proposed hybrid CNN-Vision Transformer frameworks for visually similar defects on RIAWELC [15], and developed self-configuring transformer networks for radiographic weld segmentation [16]. These developments motivate the use of modern, high-capacity architectures such as ConvNeXtV2 in the present work while showing that reported accuracy gains remain contingent on the evaluation protocol used. In particular, patch-based datasets such as RIAWELC require radiograph-level splitting (analogous to patient-level splitting in medical imaging) to prevent leakage from correlated patches derived from the same source radiograph: because neighboring patches are cropped from the same physical weld bead under the same acquisition conditions, they necessarily share structural content such as weld geometry, background texture, and sensor noise, so this correlation follows directly from how the patches were extracted rather than from a measured pixel- or feature-level correlation statistic.

2.2 Domain Shift in Industrial Radiography

Domain shift occurs when training and testing data differ in acquisition conditions, preprocessing, noise patterns, contrast, sharpness, or texture distribution. Industrial radiographic images are particularly vulnerable because imaging conditions vary across scanners, exposure settings, and digitization processes; even when a shifted radiograph remains visually plausible, subtle differences in sharpness can affect defect classification, especially for cracks and lack of penetration that rely on fine local structures. This is a specific instance of the broader dataset shift problem [17], where the bam5 case is closest to covariate shift [18] since the marginal image-feature distribution changes while defect-label semantics remain fixed, an effect that is closely tied to how partition-induced dataset shift can bias k-fold cross-validation estimates [19]. A complementary line of research develops statistical methods for detecting dataset shift by training classifiers to distinguish source from target samples [20]; the domain classifier in the present study follows this detection-by-classification principle applied at the radiograph level. Closely related to our work, recent quantification of scanner domain shift in medical X-ray imaging [21] shows that networks trained on data from one scanner consistently underperform when evaluated on data from another, providing a clinical-imaging counterpart to the industrial-radiography phenomenon we examine here.

2.3 Modern Architectures: From CNNs to Vision Transformers

The architectures applied to radiographic weld defect classification have evolved alongside the broader computer vision literature. Early work used classical CNNs (SqueezeNet, ResNet, VGG) that learn local features through stacked convolutions. The Vision Transformer (ViT) [22] showed that pure self-attention on image patches can match or exceed CNNs by modeling global relationships, and hybrid variants such as MobileViT [23] combine convolutional inductive biases with transformer blocks for accuracy-efficiency trade-offs. The ConvNeXt family [24] modernized pure CNNs with transformer-inspired design choices, and ConvNeXtV2 [25] further improved representation learning via masked autoencoder pretraining. We use ConvNeXtV2 as the primary model for its strong accuracy under a convolutional inductive bias, and also evaluate MobileViT-S, MobileNet variants, YOLO11-n-cls, and a panel of Vision Transformers (Section 4.5) to assess whether the observed domain-shift effect is consistent across architectures.

3 Materials and Methods

This section describes the seven-stage workflow shown in Fig. 1, from the original dataset through generalization testing.

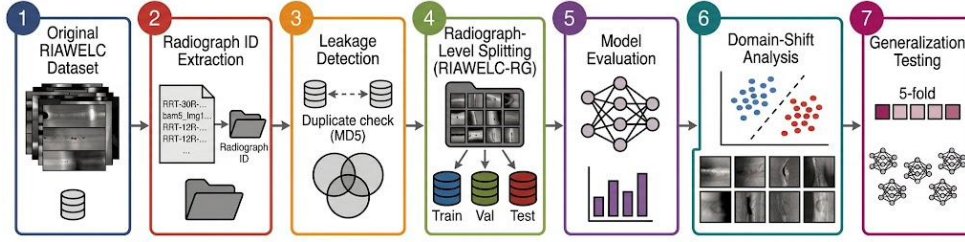


Fig. 1: Overview of the study pipeline. Stages 1-5 cover dataset acquisition, radiograph ID extraction, MD5-based leakage detection, RIAWELC-RG construction, and model training. Stage 6 applies the radiograph-level auditing protocol, combining per-radiograph performance analysis, embedding-distance-to-centroid comparison, statistical intensity profiling, and domain-classifier attribution, to bam5 as a case study. Stage 7 tests the generality of the finding via cross-validation and cross-architecture evaluation.

3.1 Dataset Description

This study uses the original RIAWELC dataset, a radiographic weld defect dataset consisting of image patches categorized into four classes: crack, lack of penetration, porosity, and no defect. The original RIAWELC images were acquired from real industrial weld radiographs digitized at 2000×8640 pixels, from which the weld bead region was cropped, extracted through windowing, resized to 224×224 pixels, and enhanced for defect visibility [7]. The original dataset organization contains 24,407 images split into training, validation, and testing subsets (Table 1). While the original dataset description reports patch counts per defect class, it does not report the number of distinct source radiographs underlying these patches. Extracting and grouping radiograph IDs from the patch filenames (Section 3.2) identifies 29 unique source radiographs underlying the 21,964 deduplicated patches in RIAWELC-RG.

Table 1: Original RIAWELC dataset organization reported in the original dataset description.

Defect Class	Total	Train	Validation	Test
Crack (CR)	7,635	4,962	1,908	765
Porosity (PO)	6,320	4,108	1,580	632
Lack of penetration (LP)	4,452	2,893	1,113	446
No defect (ND)	6,000	3,900	1,500	600
Total	24,407	15,863	6,101	2,443

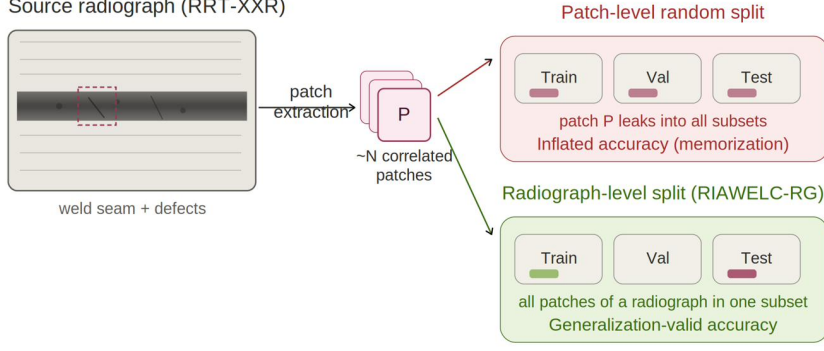


Fig. 2: Patch-level leakage risk versus radiograph-level splitting. Patches from a single source radiograph are visually correlated; random splitting allows them to appear across all subsets, inflating accuracy via memorization (top). Radiograph-level splitting keeps all patches from one source in a single subset (bottom).

3.2 Radiograph ID Extraction and Grouping

The radiograph ID was extracted from the filename prefix before the first underscore. For example, a file named RRT-30R_Img1_A80_S1_9_58.png is assigned the radiograph ID RRT-30R, while bam5_Img1_A80_S2_4_21.png is assigned the ID bam5. All patches sharing the same radiograph ID were treated as originating from the same source radiograph. This grouping enables radiograph-level splitting and per-radiograph performance analysis. Twenty-eight of the 29 identified radiographs follow a consistent RRT-XXR naming pattern; the remaining one, bam5, departs from this convention. The original dataset release does not document the meaning of these filename prefixes or the acquisition conditions behind individual radiographs, so the cause of this naming difference cannot be determined from available metadata. We therefore treat bam5 as a candidate domain-shift case to be characterized empirically from its image content rather than attributed to a presumed acquisition cause, and verify this characterization through the radiograph-level auditing protocol applied in Section 3.6 (a visual comparison of representative bam5 and RRT-domain patches is shown later in Fig. 6). The resulting radiograph IDs are used to detect leakage (Section 3.3) and to construct the radiograph-level split, with the full assignment of IDs to each subset summarized in Section 3.4 and the complete radiograph-to-subset mapping available in the released repository (see Data and Code Availability). The five radiographs assigned to the test set are RRT-15R, RRT-22R, RRT-27R, RRT-88R, and bam5, deliberately placed there as a targeted case study (Section 3.3).

3.3 Leakage Detection

Two forms of leakage were considered. First, duplicate-level leakage was checked by computing MD5 hashes of image files across training, validation, and testing subsets.

Consistent with the recent observation by Thfmm [8], this check confirms a severe problem in the original dataset organization: all 2,443 files in the original testing subset are byte-identical duplicates of files in the training subset. In other words, the entire original test set overlaps with the training data at the file level, meaning that test accuracy reported under the original split effectively measures memorization rather than generalization. Second, radiograph-level leakage was assessed by checking whether the same radiograph ID appeared in more than one subset; 28 radiograph IDs were shared between the original training and testing subsets, and 27 between training and validation as well as between validation and testing. The radiograph-level reorganization adopted in this work removes both forms of leakage, and after deduplication the resulting RIAWELC-RG dataset contains 21,964 unique patches. Table 2 summarizes the leakage levels before and after correction.

Table 2: Leakage detection summary.

Leakage Type	Before Correction	After Radiograph-Level Split
Duplicate hash overlap (Train-Test)	2,443 identical files (100% of the original test set)	0
Train-Test radiograph ID overlap	28	0
Train-Val radiograph ID overlap	27	0
Val-Test radiograph ID overlap	27	0

MD5 hashing detects only byte-identical duplicates; it does not identify near-duplicate patches that share the same underlying scene but differ slightly due to reprocessing, re-cropping, or minor acquisition variation, which would not produce identical hashes despite substantial content overlap. Radiograph-level grouping (Section 3.2) mitigates this risk for patches that share a recoverable radiograph ID, since such patches are kept together in a single subset regardless of whether their hashes match. Near-duplicates that do not share a recoverable ID, for instance if the same underlying weld were reprocessed under a different file-naming convention, cannot be ruled out from the metadata available in the original dataset release, and remain a residual risk for RIAWELC-RG as for any patch-based dataset without full acquisition provenance.

3.4 Radiograph-Level Splitting Protocol

To prevent information leakage, the dataset was reorganized using a radiograph-level split (Fig. 2). Instead of assigning individual patches randomly to training, validation, and testing subsets, radiograph IDs were first grouped. All patches belonging to the same radiograph were then assigned exclusively to one subset. This ensures that no source radiograph appears in more than one subset. Table 3 summarizes the resulting split.

Table 3: Radiograph-level split summary of RIAWELC-RG (after duplicate removal).

Split	Radiograph IDs	Patches	Classes Present	Contains bam5
Training	18	15,120	Yes	No
Validation	6	3,357	Yes	No
Testing	5	3,487	Yes	Yes

3.5 Model and Evaluation Setup

Classification models were used as diagnostic tools to evaluate the effect of radiograph-level splitting and domain shift, rather than as an exhaustive architecture benchmark; the focus is on the reliability of evaluation. ConvNeXtV2-T was used as the primary model because it achieved the strongest performance among the evaluated architectures, with ConvNeXtV2-femto, YOLO11-n-cls, MobileViT-S, and the MobileNetV1/V2/V3-Small/V4-Small family evaluated under the same protocol as supporting evidence to assess whether the bam5 effect generalizes across model capacities. Models were fully fine-tuned on the radiograph-level split, under scenarios with and without bam5. The main evaluation metrics include accuracy, macro F1-score, weighted F1-score, and per-radiograph accuracy (Table 4).

All models were trained using a standardized two-phase fine-tuning protocol: head-only training for 30 epochs at a learning rate of 1×10^{-4} , followed by full fine-tuning for 15 epochs at a learning rate of 5×10^{-6} , using CrossEntropyLoss, ImageNet-pretrained initialization (via timm where applicable), no data augmentation, and a fixed random seed (42). This protocol was applied uniformly across all evaluated architectures, including the five Vision Transformer models introduced in Section 3.7, to ensure that observed differences reflect genuine architectural responses rather than inconsistent training recipes.

Table 4: Experimental setup summary for the rigorous evaluation study.

Component	Setting
Input size	224×224
Input channels	Grayscale converted to 3-channel
Evaluation protocol	Radiograph-level split
Main models	ConvNeXtV2-T (primary), ConvNeXtV2-femto, MobileViT-S, MobileNet family, YOLO11-cls
Metrics	Accuracy, Macro F1, Weighted F1, Per-radiograph accuracy
Domain-shift analyses	Embedding distance, intensity statistics, domain classifier, feature importance

3.6 Domain-Shift Analysis of bam5

We formalize the radiograph-level auditing protocol introduced in Section 1 as a set of four complementary diagnostics comparing a candidate radiograph r against a reference domain D :

$$\mathcal{A}(r, D) = \{ P_{\text{perf}}(r), P_{\text{dist}}(r, D), P_{\text{stat}}(r, D), P_{\text{clf}}(r, D) \} \quad (1)$$

where P_{perf} is per-radiograph classification performance, P_{dist} is the embedding distance of r from the centroid of D , P_{stat} compares statistical intensity and texture features, and P_{clf} is the separability of r from D under a domain classifier. A radiograph is flagged as a domain-shift case when multiple components of $\mathcal{A}$ agree in indicating atypicality relative to D . Applied here to bam5 against the RRT domain, each component of $\mathcal{A}$ is described below.

The bam5 radiograph was analyzed using four complementary approaches. First, per-radiograph performance (P_{perf}) was computed to determine whether classification errors were concentrated on bam5. Second, feature embeddings were extracted from a trained model and used to compute the distance (P_{dist}) of bam5 from the RRT-domain centroid. The RRT-domain centroid was defined as the L2-normalized mean of L2-normalized ConvNeXtV2-T embeddings over all RRT-labeled patches pooled across the training, validation, and testing subsets, as given in Equation (2):

$$\mu_{\text{RRT}} = \text{normalize} \left(\frac{1}{N_{\text{RRT}}} \sum_{i=1}^{N_{\text{RRT}}} \text{normalize}(f_i) \right) \quad (2)$$

where $f_i \in \mathbb{R}^{768}$ denotes the raw ConvNeXtV2-T embedding of the i -th RRT patch and $\text{normalize}(x) = x/\|x\|_2$ denotes L2 normalization. Embeddings are normalized before averaging so that the resulting centroid, and the cosine and Euclidean distances computed against it, reflect differences in feature direction rather than differences in raw activation magnitude across patches. N_{RRT} is the total number of RRT patches across all three subsets combined (not the testing subset alone). Third, intensity and texture statistics (P_{stat}) were computed from raw patches. Fourth, binary domain classifiers (P_{clf}) were trained to distinguish RRT-domain patches from bam5 patches using statistical image features, with radiograph-level grouping applied on the RRT side to prevent patches from the same source radiograph from appearing in both the training and testing partitions of the classifier (Section 4.4). Together, these four components of $\mathcal{A}$ (Equation (1)) determine whether bam5 represents ordinary within-domain variation or a domain-shift radiograph.

3.7 Generalization Protocols

To verify that the bam5 performance gap and class-selective degradation observed in Sections 4.1-4.2 are not artifacts of a single partition, we conducted a 5-fold radiograph-level cross-validation using ConvNeXtV2-T. The 29 source radiographs were partitioned into five folds (Table 5) with bam5 fixed as a test radiograph in Fold 3 rather than rotated across folds; this is a targeted case-study design (Scenario A). The training and validation pool for every fold was drawn exclusively from the 28

RRT-series radiographs: in Folds 1, 2, 4, and 5, bam5 is entirely absent from training, validation, and testing, so these folds estimate typical cross-radiograph variability among RRT-only data, while Fold 3 additionally introduces bam5 as a held-out test radiograph. Across all five folds, bam5 therefore never contributes to model training; it is evaluated exclusively as unseen test data in Fold 3. This design directly evaluates the scenario in which bam5 appears only as test data; it does not evaluate bam5 appearing in the validation set, nor bam5 contributing patches to model training, both of which remain open for future work (Section 6). The remaining 28 RRT-series radiographs were distributed so that all four defect classes were represented in the training set of every fold. Within each fold, the remaining radiographs were further split into training and validation subsets at the radiograph level, the model was retrained from ImageNet-pretrained initialization with no information transferred across folds, and the same standardized two-phase fine-tuning protocol described in Section 3.5 was applied. Cross-validation outcomes are reported in Section 4.5 at both the per-fold accuracy level and the per-radiograph per-class level.

To assess generalization beyond a single architecture, we also evaluated five Vision Transformer and ViT-derived models: ViT-Tiny (vanilla) [22], DeiT-Tiny (distilled) [26], PiT-Ti (pooling-based) [27], MobileViTv2-100 (hybrid CNN-attention) [28], and XCiT-Tiny-12-p8 (cross-covariance attention) [29]. All five were evaluated under the same standardized two-phase fine-tuning protocol described in Section 3.5. These architectures span the principal design axes of modern transformer-based image classifiers (tokenization, spatial reduction, attention formulation, and convolutional inductive bias). For each model, two runs were performed with identical hyperparameters but different test compositions, one on RIAWELC-RG-nobam and one with bam5 included. Because the with-bam5 test set augments the nobam set with bam5 patches, bam5-only crack accuracy can be estimated by subtraction: applying the nobam crack accuracy to the non-bam5 portion and subtracting from the observed total. This is valid because the underlying RRT-series test patches are identical between conditions.

Table 5: Test-set composition per fold in the 5-fold radiograph-level cross-validation. Bam5 is fixed as a test radiograph in Fold 3 (Scenario A: targeted case study).

Fold	Test Radiographs	N (Test RID)
1	RRT-09R, RRT-105R, RRT-11R, RRT-13R, RRT-22R, RRT-28R, RRT-31R	7
2	RRT-102R, RRT-23R, RRT-40R, RRT-97R, RRT-98R, RRT-99R	6
3	RRT-107R, RRT-15R, RRT-24R, RRT-26R, RRT-88R, bam5	6
4	RRT-101R, RRT-12R, RRT-20R, RRT-42R, RRT-90R	5
5	RRT-27R, RRT-29R, RRT-30R, RRT-39R, RRT-94R	5

3.8 Software and Analysis Tools

Model training and feature extraction were implemented in PyTorch using the timm library. The statistical analyses reported in Sections 4.2 and 4.5 were computed using standard Python libraries (SciPy, statsmodels); the analysis scripts are included in the accompanying repository. A large language model (Claude, Anthropic) was used to assist in implementing these analyses and in drafting and revising portions of the manuscript text. All outputs were verified and edited by the authors, who take full responsibility for the content of this work.

4 Results

4.1 Impact of bam5 on Classification Performance

Under radiograph-level evaluation, model performance depended strongly on whether bam5 was included in the test set: a fully fine-tuned ConvNeXtV2-T model reached 0.9663 accuracy when bam5 was excluded (RIAWELC-RG-nobam) but dropped to 0.9277 when included (RIAWELC-RG with bam5), with the same trend observed for ConvNeXtV2-Femto and six additional lightweight architectures evaluated under the same protocol: YOLO11-n-cls [30], MobileViT-S, and the MobileNetV1/V2/V3-Small/V4-Small family (Table 6). All eight evaluated models showed a consistent accuracy gap of 0.024 to 0.074 between conditions, in the same direction across every architecture, indicating that the bam5 effect is not architecture-specific and persists across the CNN, hybrid, and detection-oriented architectures evaluated here.

For architecture a , the accuracy gap is defined as

$$\text{Gap}(a) = \text{Acc}_{\text{nobam}}(a) - \text{Acc}_{\text{withbam}}(a) \quad (3)$$

so that a positive value indicates lower accuracy when bam5 is included in the test set.

To ensure that the observed gaps reflected genuine architectural responses rather than inconsistent training recipes, the standardized two-phase fine-tuning protocol described in Section 3.5 was applied uniformly across all eight models, under which every model produced a positive gap, removing the small reversed or negligible gaps seen under earlier per-model recipes. Two caveats apply to Table 6: MobileNetV2 reached lower absolute accuracy than the other architectures and was still mildly underfitting at the end of training, though the direction of its bam5 gap matches the others; and YOLO11-n-cls was trained through the Ultralytics framework, whose scheduler and freezing strategy differ from the explicit two-phase loop used elsewhere, so its absolute accuracies carry this protocol caveat even though its bam5 gap remains positive and within range.

Table 6: Performance comparison under radiograph-level split with and without bam5. All models were fully fine-tuned; nobam = RIAWELC-RG-nobam, withbam = RIAWELC-RG with bam5. Gap is defined in Equation (3).

Model	nobam	withbam	Gap
ConvNeXtV2-T	0.9663	0.9277	0.0386
ConvNeXtV2-Femto	0.9471	0.9229	0.0242
YOLO11-n-cls	0.9124	0.8638	0.0486
MobileViT-S	0.8644	0.8208	0.0436
MobileNetV2	0.7371	0.6923	0.0448
MobileNetV3-Small	0.9024	0.8288	0.0736
MobileNetV1	0.8274	0.7591	0.0683
MobileNetV4-Small	0.8198	0.7907	0.0291

4.2 Per-Radiograph Performance

Per-radiograph evaluation showed that the performance degradation was not uniformly distributed across all radiographs. Within the radiograph-level test set, the four RRT radiographs achieved accuracies between 0.930 and 1.000, whereas bam5 achieved only 0.676 (Fig. 3). A further class-level breakdown of bam5 revealed that this degradation was concentrated almost entirely in a single class: the crack class fell to 0.380 accuracy, while lack of penetration and porosity were both classified perfectly at 1.000. Table 7 reports 95% Clopper-Pearson exact confidence intervals for each class; the crack-class interval does not overlap with either the lack-of-penetration or porosity intervals, indicating that the class-selective gap is not attributable to sampling noise in the patch counts. This class-selective pattern supports the interpretation that bam5 represents a hard-domain case in which a specific defect type, rather than the radiograph as a whole, is most affected by the domain shift.

Table 7: bam5 per-class accuracy with 95% Clopper-Pearson exact confidence intervals.

Class	Correct / Total	Accuracy	95% CI
Crack	92 / 242	0.380	[0.319, 0.445]
Lack of penetration	68 / 68	1.000	[0.947, 1.000]
Porosity	153 / 153	1.000	[0.976, 1.000]

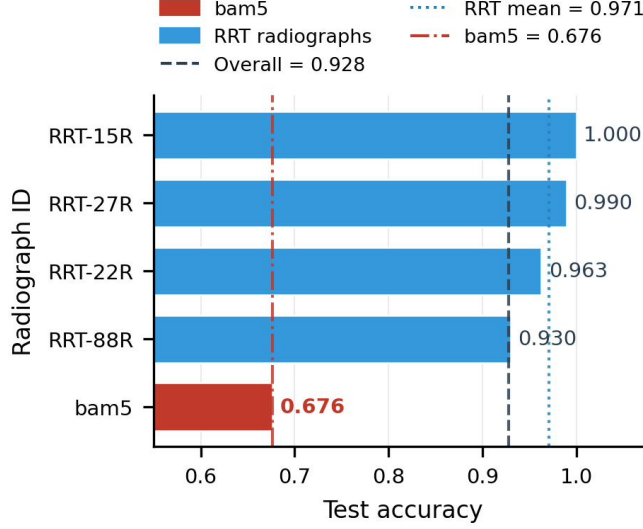


Fig. 3: Per-radiograph test accuracy (ConvNeXtV2-T, fully fine-tuned). bam5 (red) achieves 0.676, well below all four RRT radiographs (0.930–1.000). The gap of 0.270 from the RRT mean indicates a hard-domain condition rather than ordinary test variation.

4.3 Feature Embedding Distance

Feature embedding distance analysis was performed to determine whether bam5 lies far from the RRT-domain feature distribution. The underlying idea is that a fully fine-tuned model encodes each patch as a point in a high-dimensional feature space, and patches with similar visual content cluster close together; computing the distance from bam5 to the center of the RRT cluster therefore quantifies how visually atypical bam5 is relative to the RRT radiographs, independent of classification accuracy. Embeddings were extracted from a fully fine-tuned ConvNeXtV2-T model, and the distance of bam5 to the RRT-domain centroid was compared with the distances of the four individual RRT radiographs in the test set. Because only four RRT radiographs are available for this comparison, a z-score was used to express how many standard deviations bam5’s distance falls from the mean RRT-radiograph distance, and this z-score was then converted to a percentile to ease interpretation; both summarize the same comparison and are reported together for transparency rather than as independent findings. At the global level, bam5 ranked at the 67.86th percentile for both cosine and Euclidean distance (two different ways of measuring the same notion of separation in feature space, included together as a consistency check), indicating that it does not appear as an extreme outlier when all defect classes are pooled together (Table 8). This embedding-distance percentile is conceptually distinct from the per-feature outlier check in Section 4.4: the former measures overall visual similarity via

a trained model’s learned representation, while the latter checks each individual statistical image property (such as intensity or sharpness) separately for extreme values. However, a per-class analysis of the embedding distance revealed a markedly stronger pattern: bam5 lies at the 100th percentile for both the crack and lack of penetration classes, and at the 94.74th percentile for porosity. This indicates that the domain shift associated with bam5 is class-discriminative rather than a uniform distributional offset, and is most pronounced for the crack class, which also exhibits the largest drop in classification accuracy.

Table 8: Feature embedding distance of bam5 relative to the RRT-domain centroid.

Metric	Value
bam5 samples	463
Cosine distance to RRT centroid	0.2393
Cosine z-score vs RRT radiographs	0.572
Cosine percentile vs RRT radiographs	67.86
Euclidean distance to RRT centroid	0.6918
Euclidean z-score vs RRT radiographs	0.639
Euclidean percentile vs RRT radiographs	67.86

To complement these distance-based statistics with a qualitative view of the embedding space, Fig. 4 shows a t-SNE projection of the same ConvNeXtV2-T embeddings. bam5 patches (right panel, grouped by class) do not scatter randomly across the broader RRT distribution (left panel); instead they form a small number of tight, spatially distinct clusters, consistent with the per-class percentile pattern reported above rather than with a diffuse, uniformly-shifted domain. A UMAP projection of the same embeddings (not shown) produces the same qualitative clustering pattern, serving as a consistency check on the t-SNE result.

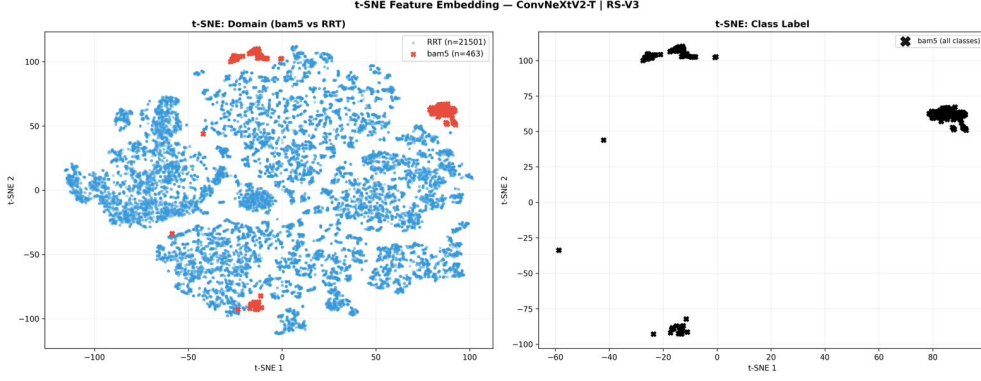


Fig. 4: t-SNE projection of ConvNeXtV2-T embeddings. Left: RRT patches (blue, $n=21,501$) versus bam5 patches (red, $n=463$). Right: the same bam5 patches, grouped by defect class. bam5 forms tight, spatially separated clusters rather than blending into the RRT distribution.

4.4 Domain Forensics: Statistical Features and Domain Classification

Radiograph-level intensity statistics do not classify bam5 as an extreme univariate outlier, as it falls within the range of normal RRT variation on a per-feature percentile basis (Table 9, Fig. 5); however, all bam5 patches fall below the RRT median sharpness, consistent with the concentrated low-sharpness peak visible in Fig. 5. A visual comparison of representative patches (Fig. 6) further confirms that bam5 patches appear uniformly blurred relative to RRT patches across all three defect classes. The combination of these features through a domain classifier, as shown next, separates bam5 from the RRT domain almost perfectly.

Table 9: Top radiograph-level shifted features for bam5 relative to the RRT domain. Pct. = extremeness percentile among RRT radiographs.

Feature	bam5 Value	RRT Mean	RRT Std	Z-score	Pct.
p95_intensity	218.663	185.050	43.063	0.781	72.74
std_intensity	200.194	170.446	41.208	0.722	70.75
max_intensity	230.255	204.116	44.411	0.589	63.56
p75_intensity	31.444	24.022	17.679	0.420	70.98
sharpness_lap_var	2.488	33.690	63.308	-0.493	1.82

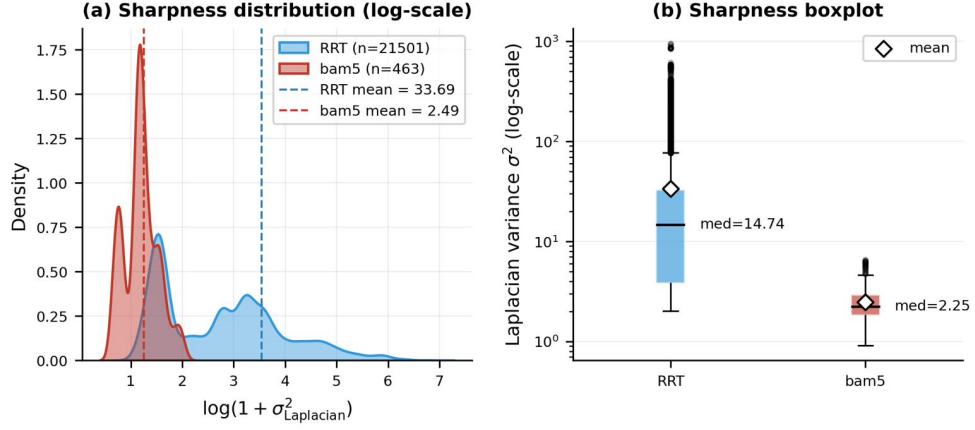


Fig. 5: Distribution of Laplacian variance (sharpness) for RRT-domain and bam5 patches. (a) Kernel density estimate on log-scale; bam5 is sharply concentrated at low values. (b) Boxplot confirms bam5 is uniformly less sharp (median 2.25 vs. RRT 14.74).

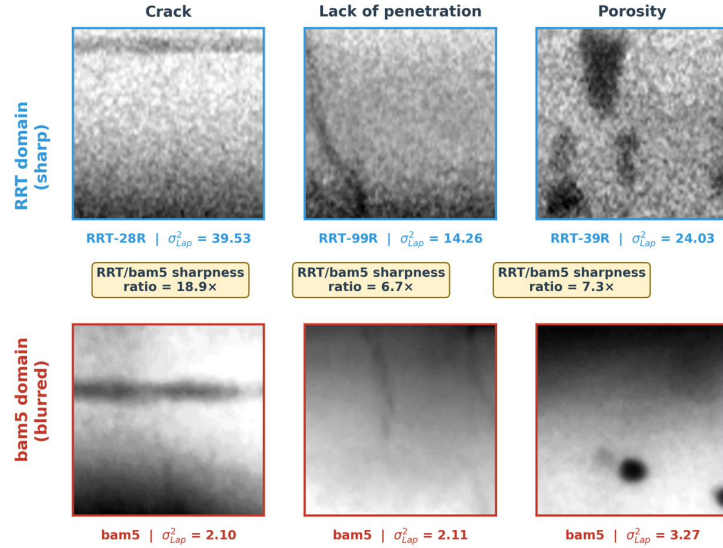


Fig. 6: Representative RRT (top, sharp) and bam5 (bottom, blurred) patches per defect class, with Laplacian sharpness annotated. Crack shows the largest sharpness ratio (18.9 $\times$), consistent with its accuracy drop to 0.380; lack of penetration (6.7 $\times$) and porosity (7.3 $\times$) remain perfectly classified.

The per-feature outlier check above considers each statistical property in isolation, so a radiograph could still be unusual through a specific combination of features even when no single feature crosses an extreme threshold. To test this possibility directly, we trained a domain classifier: a model whose only task is to distinguish bam5 patches from RRT patches using the full set of statistical image features together. If bam5 were ordinary within-domain variation, such a classifier would perform little better than chance; strong separability instead indicates a genuine, learnable difference between the two domains. The classifier is binary because the task has exactly two classes (RRT versus bam5), and we evaluated it using a Random Forest. Binary domain classifiers trained to distinguish RRT-domain patches from bam5 patches using statistical image features show that the two domains are nearly perfectly separable. To rule out radiograph-specific texture leakage, we compared three validation schemes: ungrouped, radiograph-level grouped split, and leave-one-radiograph-out (LORO); all yielded comparably high separability (Table 10). In the ungrouped setting, Random Forest achieved near-perfect results across 10 balanced repeats (Table 11), with only 3 of 186 patches misclassified (accuracy 0.984; Table 12), confirming that bam5-vs-RRT separability reflects a genuine domain-level signal rather than a classifier-evaluation artifact.

Table 10: Domain classifier validation schemes under radiograph-level controls (Random Forest). Ungrouped baseline (ROC-AUC 0.9995 ± 0.0005) is in Table 11; the LORO row reports correct-identification rate.

Validation scheme	Split unit (RRT side)	Metric	Result
Grouped (Group-ShuffleSplit)	Radiograph-level (grouped)	ROC-AUC	0.992 ± 0.011
Leave-One-Radiograph-Out	Radiograph-level (1 held out)	Correct-ID rate	0.978 ± 0.044

Table 11: Repeated balanced domain classification results for RRT vs bam5.

Metric	Logistic Regression	Random Forest
Accuracy	0.9204 ± 0.0124	0.9860 ± 0.0052
Precision	0.9009 ± 0.0132	0.9819 ± 0.0087
Recall	0.9452 ± 0.0256	0.9903 ± 0.0061
F1	0.9222 ± 0.0127	0.9861 ± 0.0052
ROC-AUC	0.9761 ± 0.0058	0.9995 ± 0.0005

Table 12: Example classification report for Random Forest domain classifier.

Domain	Precision	Recall	F1-score	Support
RRT	0.9891	0.9785	0.9838	93

Domain	Precision	Recall	F1-score	Support
bam5	0.9787	0.9892	0.9840	93
Accuracy			0.9839	186

Random Forest feature importance analysis identifies Laplacian sharpness variance as the strongest factor for distinguishing bam5 from RRT-domain radiographs, contributing approximately 50.78% of the total feature importance, far exceeding other intensity and texture features (Fig. 7). This ranking was confirmed by permutation importance ($33.0\times$ the second-ranked feature) and is not attributable to feature correlation, since `sharpness_laplacian_var` showed only low-to-moderate correlation with all other features ($r \leq 0.52$). Together, the radiograph-level sharpness statistics, near-perfect domain separability, and sharpness as the leading feature converge on the same conclusion: bam5 differs from the RRT domain primarily in image sharpness rather than in defect-relevant content.

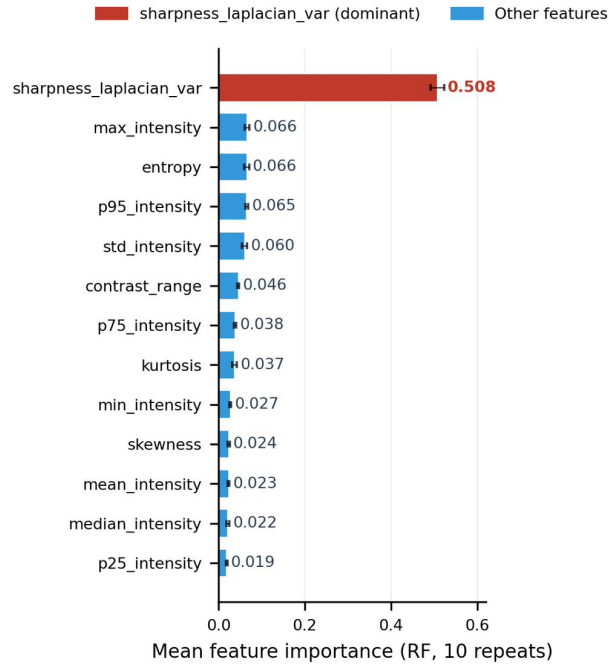


Fig. 7: Mean feature importance from a Random Forest domain classifier (10 repeats, class-balanced) for RRT vs. bam5 classification. Laplacian sharpness variance (red) dominates at 0.508, roughly $7.7\times$ the next feature (`max_intensity`, 0.066).

4.5 Generalization Beyond a Single Partition and Architecture

Following the protocol in Section 3.7, ConvNeXtV2-T was retrained from ImageNet-pretrained initialization for each of the five folds. Per-fold test accuracy ranged from 0.857 to 0.980, with a mean of 0.922 and standard deviation of 0.050 (Table 13); the four folds without bam5 scored consistently higher than Fold 3, which contains it. In folds where the test or validation set did not contain all four defect classes, class indices were explicitly aligned across splits to prevent mismatched label mappings; the reported accuracies reflect this correction.

Table 13: Per-fold test accuracy under 5-fold radiograph-level cross-validation with ConvNeXtV2-T. Bam5 is fixed as a test radiograph in Fold 3 (Scenario A).

Fold	Test Acc.	Best Val Acc.	Contains bam5	Notes
1	0.9409	0.9297	No	—
2	0.9796	0.9412	No	—
3	0.8907	0.9088	Yes	bam5 CR=0.103; RRT-107R CR=0.000
4	0.9430	0.9623	No	—
5	0.8574	0.9635	No	RRT-29R CR=0.111; RRT-30R LP=0.257
Mean	0.9223	—	—	± 0.0498 across 5 folds

The cross-validation results reproduce and extend the bam5 phenomenon. In Fold 3, the class-selective pattern reported in Section 4.2 reappears under a partition in which all training and validation radiographs differ from the primary experiment, and a second radiograph in the same fold, RRT-107R, shows the same crack-class collapse; Fold 5 surfaces two further cases, one of which (RRT-30R) is the only instance in the cross-validation where the affected class is lack of penetration rather than crack. We define a radiograph-class combination (r, c) as a class-selective failure when

$$\text{Acc}(r, c) < 0.5 \quad \text{and} \quad N(r, c) \geq 20 \quad (4)$$

The four radiograph-class combinations satisfying Equation (4) are summarized in Table 14. Three of four cases involve a crack-class collapse, consistent with fine, high-frequency crack structures being more vulnerable to acquisition-related degradation than the coarser structures of LP and porosity; the existence of RRT-30R shows the affected class is not fixed. Pooling all radiograph-class combinations evaluated across the five folds (67 combinations; 63 with $N \geq 20$) confirms this pattern formally: of the ten crack-class combinations with $N \geq 20$, three met the failure criterion, versus one of the remaining fifty-three non-crack combinations, a disproportion a one-sided Fisher’s exact test finds unlikely to arise by chance (odds ratio = 22.3, $p = 0.011$). Ranking bam5’s own crack accuracy against the other nine crack-class combinations places it second-lowest rather than uniquely extreme (RRT-107R, at 0.000, is lower still), so the cross-validation evidence supports crack-class vulnerability as a general pattern rather than singling out bam5 as a statistical outlier among failures. The key result is that bam5 is no longer the only radiograph with a failure of this magnitude. The

phenomenon recurs on independent radiographs under a partition designed to test its generality, so the bam5 case study is best interpreted as a particularly clean instance of a more general class-selective domain shift.

Table 14: Radiographs exhibiting class-selective failure (Equation (4)) under 5-fold cross-validation. Three of four cases involve the crack class.

Fold	Radiograph	Affected Class	N Patches	Accuracy
3	bam5	crack	242	0.1033
3	RRT-107R	crack	24	0.0000
5	RRT-29R	crack	63	0.1111
5	RRT-30R	LoP	381	0.2572

To assess whether this pattern generalizes beyond convolutional networks, we evaluated the panel of five Vision Transformer architectures described in Section 3.7 (ViT-Tiny, DeiT-Tiny, PiT-Ti, MobileViTv2-100, XCiT-Tiny-12-p8) on the same RIAWELC-RG-nobam and with-bam5 conditions. Table 16 reports the cross-architecture results. Across all five ViT models, the bam5 effect reproduces with the same qualitative signature as ConvNeXtV2-T: a positive accuracy gap concentrated in the crack class, with crack-to-LP misclassification as the main failure mode and no architecture exhibiting a neutral or reversed effect. The magnitude varies systematically with how strongly each architecture preserves fine-grained spatial information through its early layers: pooling-based progressive downsampling in PiT-Ti shows the largest gap, appearing to compound the loss of fine detail caused by acquisition-side blur, while distillation in DeiT-Tiny and cross-covariance attention in XCiT-Tiny retain comparatively more crack-discriminative information. Together with the eight architectures in Section 4.1, these results bring the total to thirteen architectures spanning CNNs, a detection-oriented backbone, early and recent hybrids, a vanilla ViT, a distilled transformer, a pooling-based transformer, and a cross-covariance transformer, all exhibiting a positive accuracy gap concentrated in the crack class. The class-selective domain shift documented here is therefore a property of the bam5 radiograph and its acquisition conditions, not of any specific architectural choice in the classifier.

We formally tested whether this gap reflects a systematic effect rather than architecture-specific noise by treating each of the thirteen architectures as an independent replicate and testing the null hypothesis

$$H_0 : \mu_{\text{gap}} = 0 \quad \text{vs.} \quad H_1 : \mu_{\text{gap}} > 0 \quad (5)$$

using a paired comparison between nobam and withbam accuracy (Tables 6 and 16). All thirteen architectures showed a positive gap (range 0.0188-0.0736, mean 0.0414, 95% CI [0.0315, 0.0514]). Table 15 reports three tests with different assumptions, applied to the thirteen gap values: a one-sample t-test, a Wilcoxon signed-rank test (which does not assume normality), and a distribution-free sign test based only on the direction of the gap. Because each architecture is trained and evaluated independently, this analysis avoids the patch-level correlation that complicates significance testing on individual predictions, and the agreement across all three tests rejects H_0

(Equation (5)) in favor of H_1 , indicating that the bam5 effect is a consistent property of the test data rather than an artifact of any single model.

Table 15: Paired significance tests on the nobam-withbam accuracy gap across thirteen independently trained architectures (Tables 6 and 16).

Test	Statistic	p -value
One-sample t-test	$t(12) = 9.08$	1.0×10^{-6}
Wilcoxon signed-rank	$W = 0.00$	2.4×10^{-4}
Sign test (13/13 positive)	—	1.2×10^{-4}

Table 16: Cross-architecture comparison on the bam5 effect. Estimated bam5-only crack accuracy is computed by subtraction (Section 3.7); crack→LP is the fraction of crack patches misclassified as lack of penetration (baseline nobam rate ≈ 0 for all models except a 0.6% floor for ViT-Tiny). ConvNeXtV2-T reproduces the Section 4.1 baseline.

Architecture	nobam	withbam	Est. bam5 crack	crack→LP %	Gap
ConvNeXtV2-T (Sec. 4.1)	0.9663	0.9277	—	—	0.0386
DeiT-Tiny	0.9411	0.9111	0.4586	17.9%	0.0300
MobileViTv2-100	0.8380	0.8110	0.3306	22.4%	0.0270
ViT-Tiny	0.9282	0.8884	0.2893	24.5%	0.0398
PiT-Ti	0.9587	0.9065	0.1736	27.2%	0.0522
XCiT-Tiny-12-p8	0.8998	0.8810	0.4091	25.2%	0.0188

5 Discussion

The results demonstrate that patch-level evaluation can overestimate generalization in radiographic weld defect classification. Since multiple patches originate from the same source radiograph, random patch-level splitting may allow radiograph-specific characteristics to appear in both training and testing data. A radiograph-level split is therefore necessary to assess whether a model can generalize to unseen radiographs.

The findings also show that bam5 should not simply be treated as a corrupted file or an extreme intensity outlier. Instead, bam5 is better interpreted as a hard domain-shift radiograph. It is located near the tail of the RRT feature distribution, causes severe degradation in classification performance, and is highly separable from the RRT domain using statistical image features. This combination of evidence suggests that bam5 introduces a visually coherent but shifted domain. The acquisition-side nature of this shift parallels the scanner-induced domain shift documented in medical X-ray imaging, where networks trained on data from one scanner consistently underperform when evaluated on data from another despite no change in the underlying anatomy or pathology [21]; the bam5 case is the industrial-radiography analogue of this phenomenon, with the additional property that the degradation is class-selective rather than uniform.

The dominance of Laplacian sharpness variance in the domain classifier explains the class-selective nature of the degradation. The crack class, which suffered the most severe accuracy drop on bam5 (0.380), depends on thin, elongated, low-contrast linear structures that are highly sensitive to blur; when sharpness is reduced, these fine boundaries are the first defect cues to be lost. In contrast, lack of penetration (an extended dark band along the weld centreline) and porosity (rounded high-contrast blobs) remain recognizable even when the radiograph is blurred and were both classified perfectly on bam5. This mechanism links the global feature-importance finding directly to the per-class performance. It is also worth noting that feature-space distance and classification difficulty are not perfectly aligned: the LP class lies at the 100th percentile of embedding distance from the RRT domain yet is still classified with perfect accuracy on bam5, indicating that class-level analysis is more informative than a single global distance score when characterizing domain shift.

Our findings can be informally compared with the concurrent RIAWELC-RS release by Thfmn [8], which uses a 71/16/12 patch-count partition over 15/8/6 radiographs and reports a baseline EfficientNetB0 accuracy of 81.86%; RIAWELC-RG uses an 18/6/5 radiograph partition with bam5 explicitly in the test set, reaching 0.93 with bam5 included and 0.97 when excluded under ConvNeXtV2-T. A direct comparison of these figures is uninformative because architecture, training protocol, and partitioning all differ, so to isolate partitioning from architecture we also trained an EfficientNetB0 [31] (matching Thfmn’s architecture) on our RIAWELC-RG-nobam partition, obtaining 0.8188, nearly identical to the 0.8186 reported on RIAWELC-RS. This near-identical accuracy across two different radiograph-level partitions indicates that the much lower EfficientNetB0 accuracy relative to ConvNeXtV2-T is attributable primarily to architectural capacity rather than to partitioning differences; one caveat is that bam5 sits in Thfmn’s validation set used for checkpoint selection, whereas RIAWELC-RG-nobam excludes it entirely. Both setups consistently show that leakage-free RIAWELC evaluation is much harder than the original split, and that the residual gap for our primary model is dominated by domain shift between source radiographs rather than ordinary intra-class variation.

These results have practical implications for industrial radiography datasets. Future weld defect datasets should provide source-radiograph identifiers and acquisition metadata wherever possible. Evaluation protocols should avoid random patch-level splitting and should include domain-shift analysis when source radiographs exhibit different acquisition or preprocessing characteristics. In addition, benchmark reports should include per-radiograph performance to reveal whether errors are concentrated in particular domains, rather than reporting only aggregate accuracy as in the bulk of existing weld defect classification literature [13]. The cross-validation results in Section 4.5 further suggest that the class-selective failure pattern is best understood through the lens of worst-group robustness [32]: under a fixed average accuracy, small numbers of source radiographs can carry disproportionate per-class error that is invisible to overall accuracy metrics but is consequential for safety-critical deployment. Taken together, five complementary analyses (performance drop, class-selective degradation, per-class embedding outlier behavior, near-perfect domain separability, and Laplacian sharpness as the strongest feature) converge on a single

mechanistic interpretation: bam5 differs from RRT primarily in image sharpness, and the crack class, whose visual signature is most blur-sensitive, degrades most severely. The auditing protocol responsible for this interpretation, per-radiograph performance profiling combined with embedding-distance and domain-classifier analysis, does not depend on any property specific to RIAWELC or to weld radiography; it requires only recoverable source identifiers for each patch. The same protocol can therefore be applied to other patch-based industrial NDT datasets to surface acquisition-driven domain-shift radiographs before they are mistaken for irreducible classification error.

6 Conclusion

This study introduced a radiograph-level auditing protocol for detecting acquisition-driven domain shift in patch-based industrial radiography datasets, and demonstrated it through an in-depth case study on RIAWELC. We constructed RIAWELC-RG, a leakage-free radiograph-level split, and showed that patch-based weld defect datasets require radiograph-level splitting to prevent leakage from correlated patches derived from the same source radiograph. Applying the auditing protocol to RIAWELC-RG, we identified bam5 as a hard domain-shift radiograph that strongly affects model generalization.

Although bam5 was not an extreme outlier under simple radiograph-level intensity statistics, it was located near the tail of the RRT feature distribution and was highly separable from RRT-domain radiographs using statistical-feature-based domain classifiers. Random Forest achieved an ROC-AUC of 0.9995 in distinguishing bam5 from RRT patches, and feature importance analysis revealed that Laplacian sharpness variance was the strongest separating factor. These results indicate that bam5 introduces a blur- or sharpness-related domain shift. A key limitation of this study is that the domain-shift evidence is built around a single radiograph ID, bam5, out of the 29 radiograph IDs in RIAWELC-RG. The findings presented here demonstrate that this type of blur-induced, class-selective domain shift can occur and can be detected with the proposed auditing protocol; they do not establish that such domain shift is common or representative across the broader RIAWELC radiograph population, since the remaining 28 radiograph IDs were not individually subjected to the same depth of analysis. A second limitation concerns leakage detection: MD5 hashing identifies only byte-identical duplicates and cannot rule out near-duplicate patches arising from reprocessing or minor acquisition variation that lack a shared recoverable radiograph ID (Section 3.3). Future work could further validate this finding through fully independent random splits in which the assignment of any given radiograph, including bam5, to the test set is not predetermined, through placing bam5 in the validation set or including bam5 patches during model training to test whether the class-selective effect depends on its specific role in the split, and through applying the same auditing protocol to other radiographs, and to other patch-based industrial NDT datasets, to assess how common blur-induced domain shift is beyond this one case. Without radiograph-level evaluation, per-radiograph performance reporting, and domain-shift analysis, weld defect classifiers risk overlooking acquisition-driven failures that are invisible to aggregate accuracy.

Declarations

Funding. The authors did not receive support from any organization for the submitted work.

Competing Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Data and Code Availability. The RIAWELC-RG radiograph-level split (full radiograph-to-subset ID assignments and deduplicated patches) is permanently archived with a citable DOI at <https://doi.org/10.5281/zenodo.21638849> [33]. The dataset construction and analysis code, including the domain classifier and leave-one-radiograph-out validation scripts, is available at <https://github.com/adamantoi84/riawelc-rg> and archived with a separate citable DOI at <https://doi.org/10.5281/zenodo.21640101> [34], to enable replication and fair comparison with other radiograph-level resplits of RIAWELC, including the concurrent RIAWELC-RS release by Thfmm [8].

References

- [1] Mery D, et al. GDXray: The database of X-ray images for nondestructive testing. *Journal of Nondestructive Evaluation*. 2015;34(42).
- [2] Guo W, Qu H, Liang L. WDXI: The Dataset of X-ray Image for Weld Defects. In: *Proceedings of ICNC-FSKD*; 2018. .
- [3] Tampu IE, Eklund A, Haj-Hosseini N. Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. *Scientific Data*. 2022;9(1):580. <https://doi.org/10.1038/s41597-022-01618-6>.
- [4] Yagis E, et al. Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Scientific Reports*. 2021;11:22544. <https://doi.org/10.1038/s41598-021-01681-w>.
- [5] Rouzrokh P, et al. Mitigating Bias in Radiology Machine Learning: 1. Data Handling. *Radiology: Artificial Intelligence*. 2022;4(5):e210290. <https://doi.org/10.1148/ryai.210290>.
- [6] Bai H, et al. VISION Datasets: A Benchmark for Vision-based Industrial Inspection. *arXiv preprint arXiv:230607890*. 2023;.
- [7] Vieira JP, Bauler VA, Rosa RK, Silva D. Towards a more realistic evaluation of machine learning models for bearing fault diagnosis. *arXiv preprint arXiv:250922267*. 2025;.
- [8] thfmm.: RIAWELC-RS: Welding defect classification & weakly-supervised segmentation on a leakage-free resplit of the RIAWELC radiograph dataset. GitHub. [Online]. Available: <https://github.com/thfmm/riawelc-rs>. GitHub repository, v1.0.0.

- [9] Totino B, Spagnolo F, Perri S. RIAWELC: A Novel Dataset of Radiographic Images for Automatic Weld Defects Classification. *International Journal of Electrical and Computer Engineering Research*. 2023;3(1):13–17.
- [10] Iandola FN, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. In: *International Conference on Learning Representations (ICLR)*; 2017. .
- [11] Palma-Ramírez D, et al. Deep convolutional neural network for weld defect classification in radiographic images. *Heliyon*. 2024;10(9):e30590. <https://doi.org/10.1016/j.heliyon.2024.e30590>.
- [12] López G, Duque Ramírez P, Vega E, Pizarro F, Toro J, Parra C. Weld-VGG: A VGG-Inspired Deep Learning Model for Weld Defect Classification from Radiographic Images with Visual Interpretability. *Sensors*. 2025;25(19):6183. <https://doi.org/10.3390/s25196183>.
- [13] Yang L, Jiang H. Weld defect classification in radiographic images using unified deep neural network with multi-level features. *Journal of Intelligent Manufacturing*. 2021;32(2):459–469. <https://doi.org/10.1007/s10845-020-01581-2>.
- [14] Contreras Ortiz A, Santiago RR, Hernandez DE, Lopez-Montiel M. Multiclass Evaluation of Vision Transformers for Industrial Welding Defect Detection. *Mathematical and Computational Applications*. 2025;30(2):24. <https://doi.org/10.3390/mca30020024>.
- [15] Ayele TB. A misclassification-aware explainable hybrid CNN-vision transformer framework for radiographic weld inspection. *Scientific Reports*. 2026;16(1):17134. <https://doi.org/10.1038/s41598-026-44874-x>.
- [16] Guan K, Zhou Y, Wang L, Du D. A self-configuring transformer segmentation method for welding radiographic defect detection in steel pipes. *Nondestructive Testing and Evaluation*. 2025;<https://doi.org/10.1080/10589759.2025.2491739>.
- [17] Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND, editors. *Dataset Shift in Machine Learning*. Cambridge, MA, USA: MIT Press; 2009.
- [18] Moreno-Torres JG, Raeder T, Alaíz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. *Pattern Recognition*. 2012;45(1):521–530. <https://doi.org/10.1016/j.patcog.2011.06.019>.
- [19] Moreno-Torres JG, Sáez JA, Herrera F. Study on the impact of partition-induced dataset shift on k-fold cross-validation. *IEEE Transactions on Neural Networks and Learning Systems*. 2012;23(8):1304–1312. <https://doi.org/10.1109/TNNLS.2012.2199516>.

- [20] Rabanser S, Günnemann S, Lipton ZC. Failing loudly: An empirical study of methods for detecting dataset shift. In: *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 32; 2019. .
- [21] Guo B, Lu D, Szumel G, Gui R, Wang T, Konz N, et al. The impact of scanner domain shift on deep learning performance in medical imaging: an experimental study. *International Journal of Computer Assisted Radiology and Surgery*. 2026;<https://doi.org/10.1007/s11548-026-03655-7>.
- [22] Dosovitskiy A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR)*; 2021. .
- [23] Mehta S, Rastegari M. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In: *International Conference on Learning Representations (ICLR)*; 2022. .
- [24] Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022. p. 11976–11986.
- [25] Woo S, et al. ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2023. p. 16133–16142.
- [26] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: *International Conference on Machine Learning (ICML)*; 2021. p. 10347–10357.
- [27] Heo B, Yun S, Han D, Chun S, Choe J, Oh SJ. Rethinking spatial dimensions of vision transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2021. p. 11936–11945.
- [28] Mehta S, Rastegari M. Separable self-attention for mobile vision transformers. *arXiv preprint arXiv:220602680*. 2022;.
- [29] Ali A, Touvron H, Caron M, Bojanowski P, Douze M, Joulin A, et al. XCiT: Cross-covariance image transformers. In: *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 34; 2021. p. 20014–20027.
- [30] Ultralytics.: YOLO11: Real-time object detection and classification. GitHub. [Online]. Available: <https://github.com/ultralytics/ultralytics>. GitHub repository.
- [31] Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning (ICML)*; 2019. p. 6105–6114.

- [32] Sagawa S, Koh PW, Hashimoto TB, Liang P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In: International Conference on Learning Representations (ICLR); 2020. .
- [33] Beny, Hendry, Timotius IK, Kristianto B.: RIAWELC-RG (Version 1) [Dataset]. Zenodo. [Online]. Available: <https://doi.org/10.5281/zenodo.21638849>.
- [34] Beny, Hendry, Timotius IK, Kristianto B.: riawelc-rg: Code accompanying “Revisiting RIAWELC: Radiograph-Level Auditing Reveals Data Leakage and Class-Selective Domain Shift in Weld Defect Classification”. Zenodo. [Online]. Available: <https://github.com/adamantoi84/riawelc-rg>. GitHub repository, v1.0.0.