

Quant-XAI: quantifying and mitigating explainability drift under INT8 quantization in edge-deployed visual classification (plant disease imaging case study)

Sunzil Khandaker

khandaker15-5383@diu.edu.bd

Daffodil International University <https://orcid.org/0009-0008-9422-166X>

Research Article

Keywords: Explainability drift, Model quantization, Grad-CAM, Integrated Gradients, Post-training quantization, Plant disease detection, Quantization-aware training, Explainable visual computing, Edge deployment, Model compression, Explainability stability

Posted Date: March 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9179249/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Version of Record: A version of this preprint was published at Scientific Reports on October 11th, 2026.
See the published version at <https://doi.org/10.1038/s41598-026-75299-1>.

Quant-XAI: quantifying and mitigating explainability drift under INT8 quantization in edge-deployed visual classification (plant disease imaging case study)

Sunzil Khandaker^{1,*}

¹Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh

*khandaker15-5383@diu.edu.bd

Abstract

Deep learning image classifiers are increasingly deployed to resource-constrained edge devices through INT8 post-training quantization (PTQ), including plant disease imaging. Although quantization reduces memory footprint and can improve inference efficiency, the impact on model explainability is not well comprehended. This study introduces *explainability drift*, the systematic change in saliency-based explanations between full-precision (FP32) models and quantized (INT8) models, and suggests a multi-metric evaluation suite, including top- k Intersection-over-Union, top- k Dice coefficient, and Spearman rank correlation to measure it. We apply four post-hoc explainability methods (Grad-CAM, Grad-CAM++, Integrated Gradients, and LIME) to Paddy Disease Classification data (10 classes, 10,407 rice leaf images) using three architectures of different complexity (EfficientNetV2-S, ResNet50, and MobileNetV3-Large) before and after INT8 PTQ. Our findings reveal a significant drift depending on both architecture and method. All of the models have an average top- k IoU of less than 0.17, indicating low overlap between the FP32 and INT8 top-saliency regions. Notably, results show that the depthwise-separable architecture of MobileNetV3 leads to a total collapse of Grad-CAM in INT8 with spatially uniform attribution maps (Spearman $r = 0.000$). In addition, we demonstrate that XAI-preserving quantization-aware training (QAT) with a teacher-student CAM consistency loss can improve Spearman correlation by up to 49.2%, although the results vary depending on the architecture. The work, to the best of our knowledge, provides a quantization-induced explainability drift multi-architecture benchmark and suggests practical recommendations on how to implement reliable, explainable AI in edge devices.

Keywords: Explainability drift, Model quantization, Grad-CAM, Integrated Gradients, Post-training quantization, Plant disease detection, Quantization-aware training, Explainable visual computing, Edge deployment, Model compression, Explainability stability

Introduction

The use of deep neural networks on edge systems with limited resources, such as agricultural drones, diagnostic-focused applications on smartphones, and embedded field sensors, requires aggressive model compression to ensure limited memory, latency, and power requirements. The most popular compression approach is post-training quantization (PTQ), which quantizes 32-bit floating-point (FP32) weights and activations to 8-bit integers (INT8), because it is simple, has the lowest data needs, and is supported by many hardware-accelerators[1–3]. INT8 PTQ is generally able to shrink model size by about $4\times$ with insignificant loss in accuracy on standard benchmarks, and can produce substantial inference speedups on INT8 hardware accelerators[3], which makes it an attractive route to real-life implementation.

At the same time, the use of deep learning in safety-critical and high-stakes fields, such as in medical diagnostics, autonomous navigation, and precision agriculture, has increased the need for explainable artificial intelligence (XAI). Grad-CAM[4], Grad-CAM++[5], Integrated Gradients[6], and LIME[7] are post-hoc methods of explanation that produce visual or attributional explanations of model predictions and allow domain experts to ensure that models are attentive to physiologically meaningful attributes. Saliency maps, which are maps that show where the lesions are and not the foliage surrounding them, are used in detecting disease in plants to enhance confidence that the classification is based on disease-related features[8]. The

EfficientNetV2[9], ResNet[10], and MobileNetV3[11] among the architectures that are accurate and deployment efficient, are three different design philosophies, compound scaling, residual learning, and depthwise-separable convolutions, that allow conducting a principled exploration of the interaction between architectural choices and explainability changes induced by quantization.

Although both quantization and XAI studies matured similarly and at the same pace, their interplay is as poorly studied as it is crucial. Quantization changes the internal representations of a network, rounding weights, clipping activations, and merging fine-grained feature differences, but the post-hoc explanation implications of the downstream changes are virtually uninvestigated. When quantization modifies the internal computations of a model without modifying anything it predicts, without leaving a trace, practitioners may find themselves putting convincingly-looking but misleading explanations in safety-critical edge devices. The concern is especially evident in precision agriculture, where agronomists can use saliency maps to ensure that a classifier is paying attention to disease symptoms and not deciding based on background texture, and in medical imaging, where gradient-based attributions can be used to make clinical decisions. Although we study plant disease imaging here, the same risk applies broadly to edge-deployed visual classification systems that rely on saliency maps for accountability.

To explain the systematic change in saliency-based explanations, which occurs when a model is INT8 quantized, we coin the word **explainability drift**. To measure this drift, we propose a three-metric evaluation suite: top- k Intersection-over-Union (IoU), top- k Dice coefficient, and Spearman rank correlation r . Our three contributions in this study:

- **Explainability drift measurements.** We define the explainability drift and suggest a standardized protocol of multi-metric evaluation that will allow us to compare saliency preservation in a quantized and reproducible manner.
- **Extensive empirical point of reference.** Our systemic comparison of four mainstream XAI algorithms (Grad-CAM, Grad-CAM++, Integrated Gradients, and LIME) on three architectures of varying complexity (EfficientNetV2-S, ResNet50, and MobileNetV3-Large) on the Paddy Disease Classification dataset with 10 classes before and after INT8 quantization (ONNX Runtime PTQ for deployment; matched PyTorch fake-quantization for gradient-accessible XAI drift measurement), provides, to our knowledge, a multi-architecture, multi-method comparison of quantization-induced explainability drift, demonstrated on plant disease imaging.
- **Mitigation analysis.** We quantize training with XAI-preserving quantization-aware training (QAT) with a teacher–student CAM consistency loss, examine the vulnerability of each class, and give recommendations on useful deployment guidelines to trustworthy edge AI.

The organization of the rest of the paper is as follows. The Literature Review provides a survey of previous studies on quantization and XAI, their intersection, and the remaining gaps. The Methods section provides the description of the dataset, model architectures, the training protocol, the quantization strategy, XAI implementations, and the drift measure protocols. The Results section reveals our empirical findings. The Discussion section interprets the results, examines limitations, and suggests directions for the future, followed by the Conclusion.

Literature Review

Quantization for efficient inference

Neural network quantization provides a lower quality of numerical representation of weights and activations in order to reduce compute and memory costs. Gholami et al.[1] and Nagel et al.[2] present extensive taxonomies of post-training quantization (PTQ), quantization-aware training (QAT), and mixed-precision strategies. Regarding edge deployment, Jacob et al.[3] reported integer-arithmetic-only inference with almost no loss of accuracy, as proven to be feasible in practice with INT8 inference on mobile hardware. Much more recently, Qin et al.[12] suggested information-entropy-driven bit-width allocation of mixed-precision quantization, sliding window smoothed entropy with adaptive assignment of layer-wise bit width on ResNet architectures using CIFAR-10/100.

Biswas et al.[13] proposed a regularization-based QAT model that concurrently tackles quantization, stuck-at faults, and device variability of ANNs and SNNs. Lu et al.[14] extended QAT to the Segment Anything Model for the medical image segmentation task and revealed the INT8 viability on eleven different imaging modalities. The first survey of zero-shot quantization (ZSQ) was introduced by Kim et al.[15], who compared more than 25 different methods and divided them into synthesis-free, generator-based, and noise-optimization ones. The PRISMA systematic review of quantization-optimized lightweight Transformer architecture fruit detector on edge devices, by Maulana and Ramasamy[16], found that, at aggressive bit widths, QAT is always superior to PTQ, yet identified the existence of critical hardware–software gaps.

Explainable AI for plant disease detection

Plant pathology based on deep learning has become explainable AI. Nagasubramanian et al.[8] were the first to describe explainable plant disease detection with 3D deep learning on hyperspectral imagery, saliency maps showing near-infrared wavelengths to be most significantly correlated with soybean charcoal rot, which is in line with known plant physiology. Nigar et al.[17] obtained 99.69% accuracy using EfficientNet-B0 with a 38-class subset of PlantVillage and implemented LIME explanations in an application that can be downloaded on a mobile device. Porfírio et al.[18] compared Vision Transformer B-32 and ResNet-50 on PlantVillage with LIME, Grad-CAM, and occlusion-based attribution and reported that high-accuracy models generate more focused and confident saliency maps.

Gani et al.[19] suggested PapayaNet, an attention-based CNN to classify papaya disease based on Grad-CAM and LIME (98.79% accuracy). Nurullah et al.[20] employed multi-label tomato disease detection with Vision Transformers using Grad-CAM, Integrated Gradients, and LIME, and demonstrated that the methods developed diverged in explanations. Preotee et al.[21] compared five variants of CAM (GradCAM, GradCAM++, LayerCAM, ScoreCAM, and FasterScoreCAM) on six architectures of Bangladeshi leaf diseases systematically. Other works are SHAP-based explanations of EfficientNet-B0[22], CNN-ViT hybrids with Grad-CAM on mulberry leaf disease (95.60% accuracy)[23], PlantXViT (a small ViT-CNN hybrid with $\sim 850K$ parameters)[24], CBAM-enhanced classification using Grad-CAM[25], and entropy-based high-level features in early plant stress detection[26]. Kusuma and Rajkumar[27] present a general overview of AI methods to detect and classify plant leaf diseases and determine gaps in the widespread use of XAI.

Intersection of XAI and quantization

There is a small but growing literature at the crossroads of explainability and quantization, but the connection has been analyzed nearly entirely in the direction of how to use XAI to enhance quantization, as opposed to quantifying the impact of quantization on the outputs of XAI.

Becking et al.[28] presented ECQx, an entropy-constrained quantization model inspired by Layerwise Relevance Propagation (LRP). With LRP heatmaps to rank the weights with the most relevance to explain predictions correctly, ECQx found up to $103\times$ compression on speech and vision tasks without compromising the quality of explanations. Ranjan and Savakis came up with two complementary versions of Vision Transformer, where LRP-QViT[29] allocates mixed-precision bit widths using LRP-derived scores of importance in each layer, and Mix-QViT[30] builds upon this with a dual-criteria scheme that uses LRP scores of importance and quantization sensitivity analysis via Integer Quadratic Programming.

In a counterintuitive result, Tashiro and Awano[31] found that activation quantization can also improve explainability, when important information is concentrated into sign bits, and when sign-bit-limited attention is used, it can achieve $1.14\times$ higher AUC on deletion metrics with 10-bit quantization of self-driving data. This implies that the quantization–XAI relationship is non-trivially directional and a bidirectional relationship.

Maleki et al.[32] trained Saliency-Guided Training (SGT) on MNIST and CIFAR-10 with ResNet-20 in the saliency-guided quantization paradigm, demonstrating that quantized models trained with saliency awareness have slightly higher accuracy and steeper feature-masking sensitivity curves. Rezabeyk et al.[33] generalized this to FashionMNIST and found that smaller bit widths (2 bits) give stronger drops in both accuracy and saliency-map clarity, quantitatively relating the bit width to the quality of the explanation. At the conference scale, SYNQ[34] proposed a Grad-CAM alignment loss that uses explicit patterns of loss between the full-precision teacher and the quantized student, which penalizes divergence in the pattern of class-activation-map between the full-precision teacher and the quantized student, and showed an accuracy improvement up to 1.74%p. Xiao et al.[35] suggested TACQ to the LLMs, wherein saliency-based

interpretability could be used to define task-critical circuits in the model that are more precise during mixed-precision allocation.

Similar XAI-guided compression work outside quantization has been done by Otsuki et al.[36], who applied LRP to post-training ResNet skip connections with conserved value, and Malik et al.[37], who applied DL-Backtrace, LRP, and Integrated Gradients to guide post-training pruning on medical images.

Cross-domain XAI applications

In addition to plant disease and quantization, XAI is used for a wide range of fields, demonstrating its relevance to the universal value of model trustworthiness. Hasan et al.[38] used Grad-CAM and LIME to classify 3D MRI brain tumors (97–99% accuracy), and Wang et al.[39] used SHAP and attention-based heat maps to diagnose coronary microvascular dysfunction. Bennett et al.[40] introduced an interdisciplinary expert panel model of bias in XAI outputs and showed that social science practices could be used to uncover selection bias and confounding in saliency accounts. De Fazio et al.[41] applied rule extraction (SkoPeRules) in medical diagnosis in clinical decision support, whereas McDonnell et al.[42] applied multi-method XAI (SHAP, LIME, Dalex) in insurance claim prediction. Yoshida et al.[43] in environmental science employed SHAP with LightGBM in the estimation of dose-rates in Fukushima, and Bañales et al.[44] in the estimation of the flexibility of household energy-demand using TreeSHAP. Guo et al.[45] surveyed explainable deep reinforcement learning on 6G network slicing. All these cross-domain studies highlight the importance of the fact that the quality of explanations is a universal issue, and it is not specific to this or that application, and the effect of deployment-oriented changes like quantization on the quality of the outputs of XAI is in urgent need of understanding.

Research gap

The above-mentioned survey indicates a rather obvious asymmetry: current research at the quantization–XAI interface utilizes XAI almost solely to enhance or direct quantization (ECQx[28], LRP-QViT[29], Mix-QViT[30], SYNQ[34], TACQ[35], SGT+PACT[32, 33]). Few studies have quantitatively assessed the drift in quality of post-hoc explanation across a variety of XAI methods and across a variety of architectures on a real-world plant disease dataset using formal quantitative measures under standard INT8 PTQ, the most widely used quantization scheme deployed in production edge deployment. This gap is the main driving force of the current study.

Table 1. Comparative analysis of related quantization–XAI works and novelty gaps addressed.

Work	Approach	Finding	Gap
ECQx[28]	LRP-guided PTQ	103× compression	Single domain
LRP-QViT, Mix-QViT[29, 30]	LRP-bit allocation	Mixed-precision quantization	ViTs only
SGT + Rezabeyk[32, 33]	Saliency-guided PTQ	Clarity retention	Toy datasets
SYNQ[34]	Grad-CAM alignment loss	align-1.74%p gain accuracy	Single method XAI
Nurullah al.[20]	Multi-XAI	Method divergence	No quantization
This work	4 XAI × 3 CNN × INT8 PTQ	Systematic drift quantification	Fills asymmetry

Methods

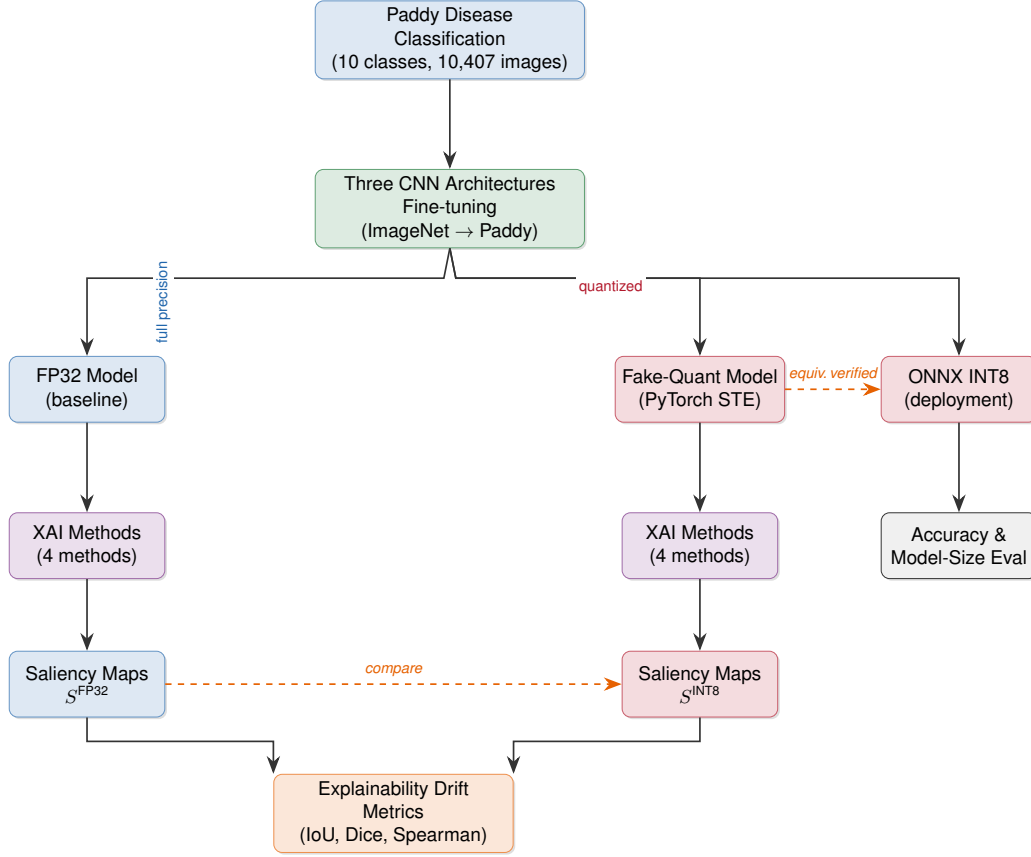


Figure 1. Quant-XAI experimental pipeline. Each trained model yields two quantized variants: an ONNX Runtime INT8 model for deployment evaluation (accuracy and model size), and a PyTorch fake-quantized model that simulates identical INT8 arithmetic while retaining gradient access for XAI computation (prediction equivalence verified between the two). Four post-hoc XAI methods generate saliency maps from both FP32 and fake-quantized models on identical evaluation images, and explainability drift is quantified using three complementary metrics.

Figure 1 illustrates the complete Quant-XAI experimental pipeline. Each component—dataset, model training, quantization, XAI computation, and drift measurement—is detailed in the subsections below.

Dataset and preprocessing

The Paddy Disease Classification dataset[46], a publicly available benchmark for rice leaf disease detection, was selected for this study. The dataset comprises 10,407 high-resolution RGB photographs of rice (*Oryza sativa*) paddy leaves, captured under field conditions and annotated by agricultural experts into 10 mutually exclusive classes reflecting distinct disease categories and healthy baseline (Table 2).

The diversity of lesion morphologies across these classes—ranging from localised, high-contrast lesions (blast, brown spot) to diffuse, low-contrast symptoms (downy mildew, tungro)—provides a natural testbed for examining how disease-specific visual characteristics modulate quantization-induced explainability drift.

Images were resized to 224×224 pixels and normalised using ImageNet channel statistics:

$$\mu = [0.485, 0.456, 0.406], \quad \sigma = [0.229, 0.224, 0.225].$$

The dataset was split into training (80%) and validation (20%) partitions via stratified sampling to preserve

Table 2. Paddy Disease Classification dataset: class taxonomy and representative lesion patterns relevant to XAI saliency analysis.

Class ID	Disease	Lesion pattern
0	Bacterial leaf blight	Linear water-soaked lesions
1	Bacterial leaf streak	Narrow interveinal streaks
2	Bacterial panicle blight	Panicle discolouration
3	Blast	Diamond-shaped necrotic lesions
4	Brown spot	Circular brown lesions with halos
5	Dead heart	Central shoot dieback
6	Downy mildew	Diffuse chlorotic patches
7	Hispa	Tunnelled transparent patches
8	Normal (healthy)	Uniform green, no lesions
9	Tungro	Orange-yellow discolouration

per-class proportions. Standard augmentations—random resized crop, horizontal flip, and normalisation—were applied during training.

Model architectures

Three convolutional neural network architectures of varying complexity and design philosophy were selected to investigate how architectural choices interact with quantization-induced explainability drift:

1. **EfficientNetV2-S[9]:** compound-scaled architecture combining Fused-MBConv and standard MBConv blocks with squeeze-and-excitation (SE) attention, representing modern efficient design.
2. **ResNet50[10]:** standard residual architecture with skip connections, representing the classical deep learning paradigm with well-understood gradient flow properties.
3. **MobileNetV3-Large[11]:** depthwise-separable convolutions with hard-swish activations and SE blocks, representing ultra-efficient mobile-deployment architectures.

All three models were initialised with ImageNet-1K pre-trained weights from the `timm` library[47].

Training protocol

Each architecture was fine-tuned for 12 epochs using the AdamW optimiser (learning rate = 1×10^{-3} , weight decay = 1×10^{-4}), cosine annealing schedule with 3 warm-up epochs, batch size 48, and automatic mixed-precision (AMP) training. Cross-entropy loss was used for all models. The best checkpoint was selected based on validation F1 score. All training was conducted on a single NVIDIA T4 GPU (Google Colab).

INT8 post-training quantization

Each trained FP32 model was exported to ONNX format (opset 17) for deployment evaluation. INT8 post-training quantization (PTQ) was performed via ONNX Runtime’s quantization toolkit (Figure 2) using static quantization with Quantize-Dequantize (QDQ) node insertion and MinMax calibration on a representative subset of training images. The pipeline includes an automatic fallback to dynamic quantization for any layer with unsupported operators; in practice, all three architectures were successfully quantized via the static path (Table 4). For XAI evaluation requiring gradient access, an equivalent fake-quantized model was constructed in PyTorch (see below), configured with identical quantization parameters to match the ONNX Runtime INT8 deployment model.

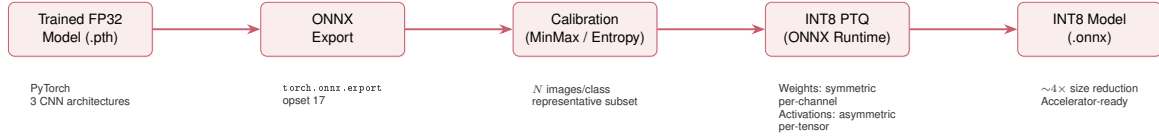


Figure 2. INT8 post-training quantization pipeline. The trained FP32 model is exported to ONNX format, calibrated with a representative image subset, and quantized to INT8 via the ONNX Runtime quantization toolkit.

XAI method implementations

Four post-hoc explanation methods were applied to identical evaluation images for both FP32 and fake-quantized models (Figure 3):

1. **Grad-CAM[4]:** gradient-weighted class activation maps from the last convolutional layer, implemented via custom PyTorch hooks for FP32 models.
2. **Grad-CAM++[5]:** pixel-wise weighting using second- and third-order derivatives; same target layer as Grad-CAM.
3. **Integrated Gradients[6]:** attribution via path integral from a black baseline (0) to the input image; 50 interpolation steps.
4. **LIME[7]:** Local Interpretable Model-agnostic Explanations with 256 perturbed samples per image and quickshift superpixel segmentation.

To ensure a methodologically valid comparison, all four XAI methods must be applied identically to both FP32 and quantized models. Since the ONNX Runtime inference engine used for INT8 deployment does not expose gradient computation, we employ PyTorch’s fake-quantization framework to simulate INT8 post-training quantization while retaining full gradient flow. Specifically, the trained FP32 model is instrumented with `FakeQuantize` modules on all weight and activation tensors, configured to match the ONNX Runtime PTQ settings (symmetric per-channel weight quantization, asymmetric per-tensor activation quantization, MinMax calibration). During the forward pass, these modules simulate INT8 rounding and clamping; during backpropagation, gradients propagate through the straight-through estimator (STE)[3], enabling Grad-CAM, Grad-CAM++, and Integrated Gradients to be computed using identical implementations on both FP32 and fake-quantized models. Drift in gradient-based attributions therefore arises from quantization-induced changes to intermediate activations and weight values, not from methodological differences in attribution computation. LIME, being inherently model-agnostic, is applied identically to both model variants without modification.

Prediction equivalence verification. To confirm that the fake-quantized PyTorch model faithfully represents the deployed ONNX Runtime INT8 model, we evaluated both on the full validation set (2,082 images) and observed high top-1 prediction agreement for all three architectures, confirming that the simulated quantization closely replicates the arithmetic behaviour of the deployed INT8 model. Exact per-architecture agreement rates and logit differences are reported in the released evaluation script.

Explainability drift metrics

For each evaluation image i , the FP32 saliency map S_i^{FP32} and fake-quantized saliency map S_i^{INT8} (both min-max normalised to $[0, 1]$) were compared using three complementary metrics (Figure 4):

Throughout this study, we fix $k = 15\%$ (i.e., the top 15% most salient pixels); this threshold was selected to capture diagnostically relevant lesion regions, which typically occupy 10–20% of the image area in the Paddy Disease dataset.

Top- k Intersection-over-Union (IoU). Both maps were binarised at the top- $k\%$ threshold, retaining only the most salient $k\%$ of pixels. IoU was computed between the resulting binary masks:

$$\text{IoU}_k = \frac{|B_k^{\text{FP32}} \cap B_k^{\text{INT8}}|}{|B_k^{\text{FP32}} \cup B_k^{\text{INT8}}|} \quad (1)$$

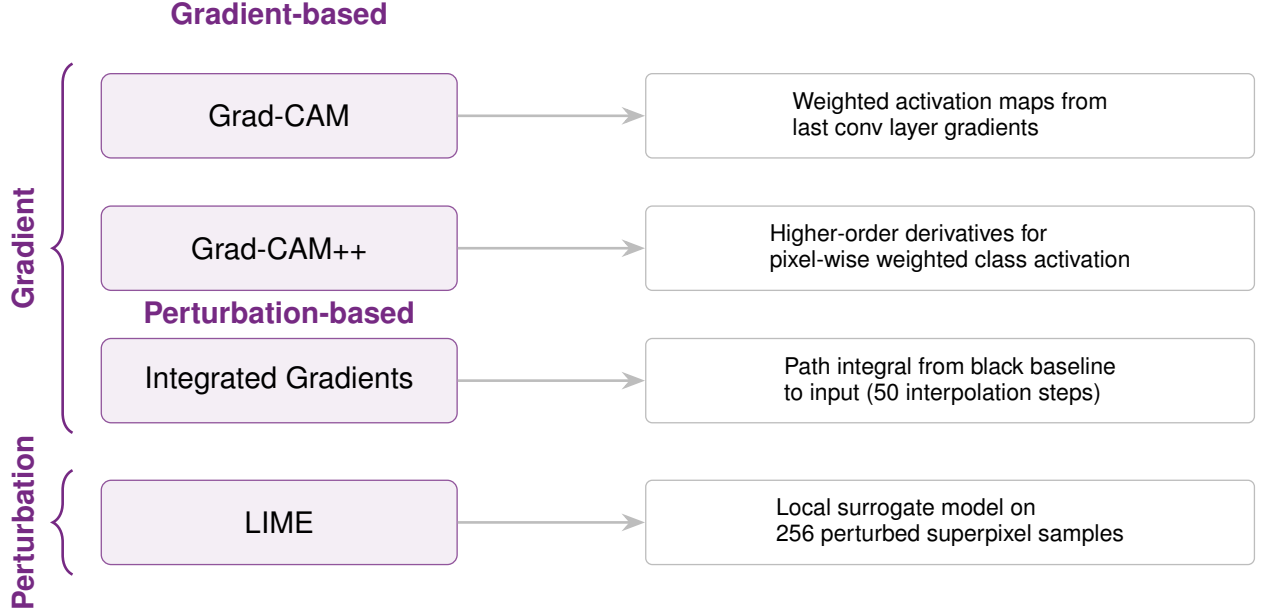


Figure 3. Four post-hoc XAI methods evaluated in this study. Three gradient-based methods (Grad-CAM, Grad-CAM++, Integrated Gradients) and one perturbation-based method (LIME) are applied identically to both FP32 and INT8 model variants.

This metric captures spatial overlap of the highest-activation regions and is sensitive to localised shifts in saliency focus.

Top- k Dice Coefficient. The Dice coefficient (F1 score of spatial overlap) is computed on the same binarised masks:

$$\text{Dice}_k = \frac{2|B_k^{\text{FP32}} \cap B_k^{\text{INT8}}|}{|B_k^{\text{FP32}}| + |B_k^{\text{INT8}}|} \quad (2)$$

Dice complements IoU by providing a less penalty-heavy measure of overlap, particularly when the salient regions are small relative to the full map.

Spearman Rank Correlation (ρ). The Spearman rank correlation between the flattened, continuous-valued saliency maps quantifies whether the *relative ordering* of pixel attributions is preserved under quantization:

$$\rho = 1 - \frac{6 \sum_{j=1}^N d_j^2}{N(N^2 - 1)} \quad (3)$$

where d_j is the difference in ranks of pixel j between the two maps. A Spearman correlation of $\rho = 1$ indicates perfect rank preservation; $\rho = 0$ indicates no monotonic relationship (i.e., the INT8 map assigns effectively random relative importance to pixels). This metric is critical for detecting scenarios where spatial overlap may be preserved by chance but the underlying attribution structure has collapsed.

XAI-preserving quantization-aware training

To mitigate explainability drift, we employ quantization-aware training (QAT) with a teacher–student CAM consistency loss. The FP32-trained model serves as the teacher, and a quantization-aware copy (with fake-quantize nodes inserted on weights and activations) serves as the student. The combined loss is:

$$\mathcal{L}_{\text{QAT}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \text{MSE}(\text{CAM}_{\text{teacher}}, \text{CAM}_{\text{student}}) \quad (4)$$

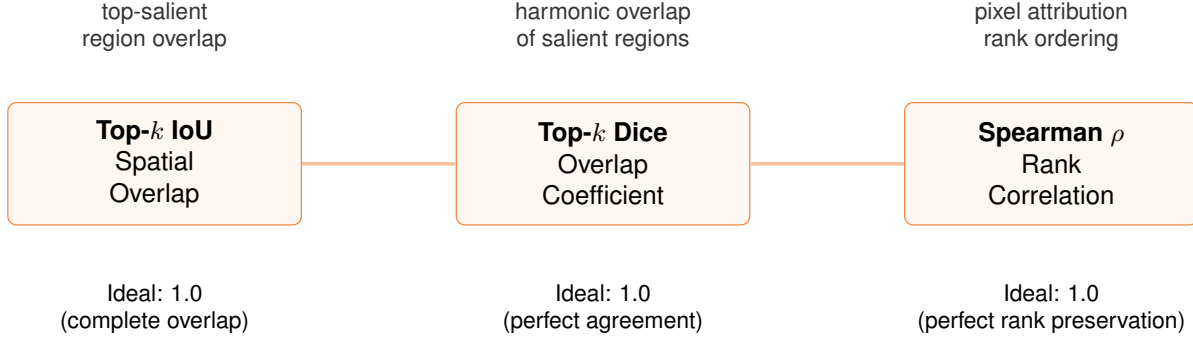


Figure 4. Explainability drift metrics used to quantify saliency-map divergence. Top- k IoU and top- k Dice measure spatial overlap of the most salient regions; Spearman ρ quantifies rank-order preservation of pixel attributions.

where $\mathcal{L}_{CE}$ is the standard cross-entropy classification loss, $CAM_{teacher}$ and $CAM_{student}$ are the Grad-CAM activation maps extracted from the last convolutional layer of the teacher and student networks respectively, MSE denotes mean squared error, and $\lambda = 0.5$ controls the trade-off between classification accuracy and explainability preservation. The student model was fine-tuned for 5 epochs with a reduced learning rate (1×10^{-4}), after which it was re-exported to ONNX and re-quantized to INT8 for drift re-evaluation.

Statistical analysis

All metrics are reported as mean $\pm$ standard deviation across the evaluation set. XAI drift was evaluated on 20 randomly sampled validation images (two per class, ensuring balanced class coverage) for each of the three architectures and four XAI methods, yielding 240 individual drift measurements per metric. We acknowledge that this sample size, while sufficient to identify major drift patterns and architectural failure modes, limits statistical power for detecting small effect sizes; future work should evaluate on the full validation set.

To control for the confounding effect of prediction changes on explanation drift, we report results both over all evaluation images and over the *matched-prediction subset*—images where FP32 and fake-quantized models agree on the predicted class—ensuring that observed drift reflects quantization’s effect on the explanation mechanism rather than a change in the explained decision.

The coefficient of variation ($CV = \sigma/\mu$) is used to assess intra-method stability. Improvement percentages for QAT mitigation are computed as relative change from baseline: $\Delta = (m_{mitigated} - m_{baseline}) / |m_{baseline}| \times 100\%$. Spearman rank correlation was computed using SciPy’s `spearmanr` implementation; for saliency maps with zero variance (constant-valued), where all pixels receive tied ranks, ρ is reported as 0.000—indicating complete attribution collapse rather than absence of correlation. All analyses used NumPy, pandas, and SciPy.

Results

This section presents the classification performance, prediction stability, and explainability drift results for the three architectures evaluated on the Paddy Disease Classification dataset. All drift results are reported over 20 randomly sampled validation images (two per class) and four XAI methods.

FP32 baseline classification performance

Table 3 reports the best validation performance achieved by each architecture during FP32 training. All three models attained strong classification performance, with validation F1 scores ranging from 0.932 (ResNet50) to 0.965 (EfficientNetV2-S). EfficientNetV2-S and MobileNetV3-Large achieved comparable balanced accuracy (~ 0.970), substantially outperforming ResNet50 (0.943). Convergence profiles differed across architectures: EfficientNetV2-S and MobileNetV3-Large converged rapidly, exceeding 94% validation F1 by epoch 6,

whereas ResNet50 exhibited slower convergence and continued improving through the final epoch. Training curves are shown in Figure 5.

Table 3. FP32 baseline classification performance across three architectures on the Paddy Disease Classification dataset (10 classes, 12 training epochs).

Architecture	Val F1	Val Balanced Acc.	Best Epoch
EfficientNetV2-S	0.9650	0.9693	11
MobileNetV3-Large	0.9616	0.9697	10
ResNet50	0.9320	0.9432	12

INT8 quantization configuration

Table 4 reports the quantization configuration successfully applied to each architecture. All three models were quantized via the static QDQ path without requiring dynamic fallback.

Table 4. INT8 post-training quantization configuration per architecture. All three models used static QDQ quantization with MinMax calibration; no dynamic fallback was required.

Architecture	Mode	Weight	Activation	Calib.	INT8 Size
EfficientNetV2-S	Static QDQ	Sym. per-ch.	Asym. per-tensor	MinMax	23.3 MB
ResNet50	Static QDQ	Sym. per-ch.	Asym. per-tensor	MinMax	24.1 MB
MobileNetV3-Large	Static QDQ	Sym. per-ch.	Asym. per-tensor	MinMax	4.9 MB

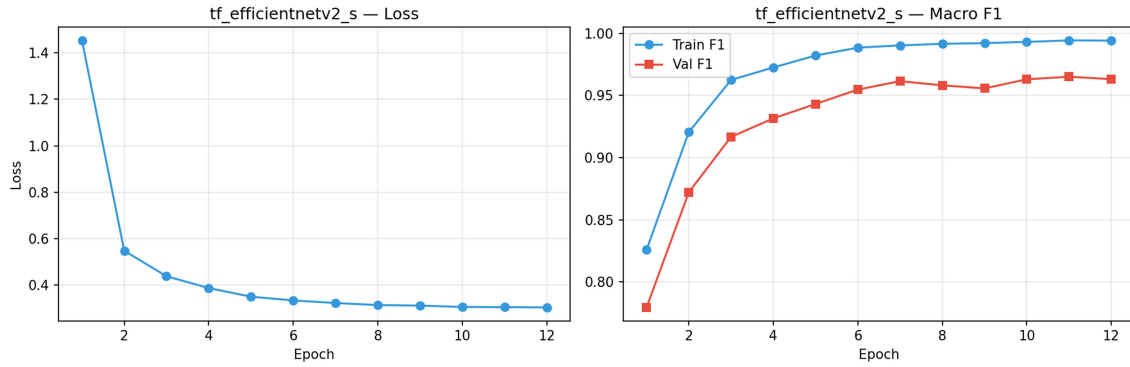
INT8 quantization impact on predictions

Beyond aggregate accuracy, INT8 quantization induced notable changes in per-image prediction labels. Table 5 presents the prediction agreement rates between FP32 and INT8 models on the 20 evaluation images.

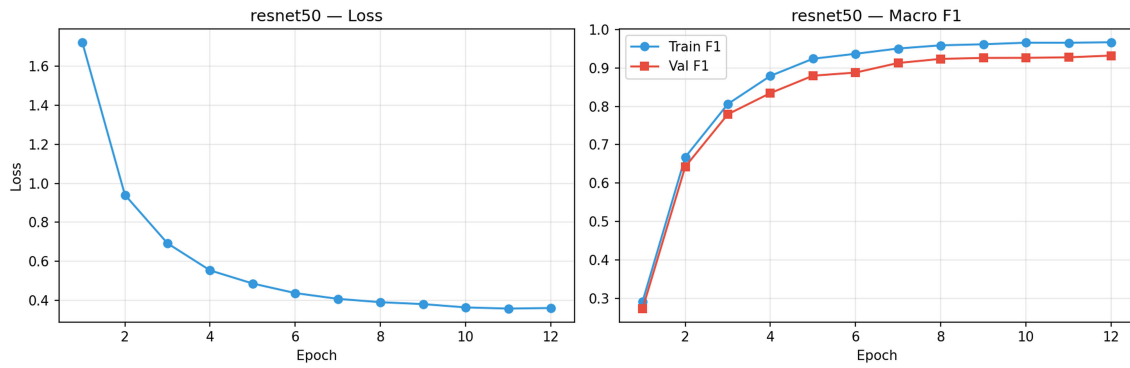
Table 5. FP32 vs. INT8 prediction agreement on 20 evaluation images. Agreement denotes identical predicted class between FP32 and fake-quantized models.

Architecture	Agreement	Percentage
MobileNetV3-Large	18/20	90.0%
ResNet50	15/20	75.0%
EfficientNetV2-S	12/20	60.0%

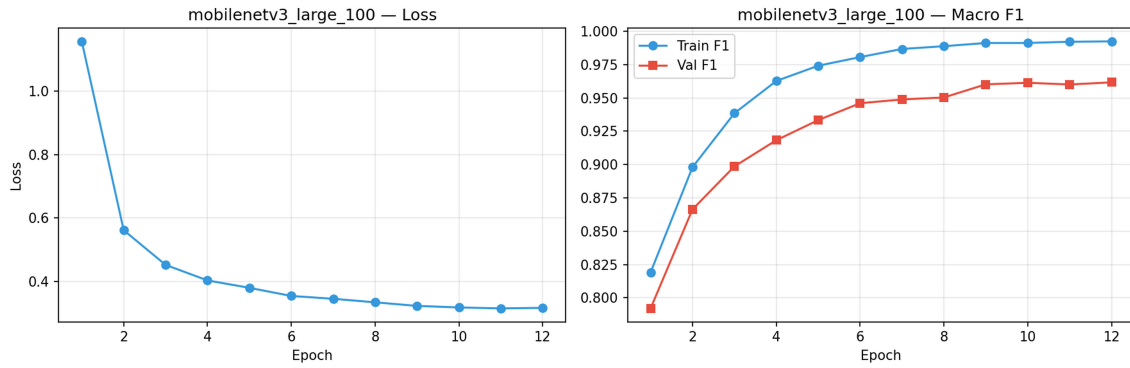
MobileNetV3-Large exhibited the highest prediction stability (90.0%), suggesting that its depthwise-separable architecture and squeeze-and-excitation blocks produce representations that are relatively robust to quantization noise at the output level. EfficientNetV2-S showed the lowest agreement (60.0%)—a counterintuitive finding given its superior FP32 accuracy—which may reflect the sensitivity of its compound-scaled architecture, with tightly coupled width, depth, and resolution scaling, to precision reduction. These prediction discrepancies are critical context for XAI evaluation, since the explanations generated for different predicted classes are fundamentally different, confounding drift measurement with prediction change. To isolate the effect of quantization on the explanation mechanism itself, all subsequent drift analyses are reported both over the full 20-image set and over the matched-prediction subset (12, 15, and 18 images



(a) EfficientNetV2-S



(b) ResNet50



(c) MobileNetV3-Large

Figure 5. Training convergence curves for (a) EfficientNetV2-S, (b) ResNet50, and (c) MobileNetV3-Large. Validation F1 score and training loss are shown across 12 epochs.

for EfficientNetV2-S, ResNet50, and MobileNetV3-Large, respectively). Supplementary Table S1 reports drift metrics restricted to the matched-prediction subset; the trends remain qualitatively unchanged from the full-set results.

Explainability drift across XAI methods and architectures

Table 6 presents the per-model, per-method baseline drift metrics, and Figure 6 visualises the cross-architecture comparison.

Table 6. Explainability drift under INT8 PTQ (mean $\pm$ std over 20 evaluation images). Higher values indicate better preservation. Best per metric in **bold**.

Architecture	XAI Method	Top- k IoU	Top- k Dice	Spearman ρ
EfficientNetV2-S	Grad-CAM	0.105 ± 0.062	0.185 ± 0.094	0.010 ± 0.118
	Grad-CAM++	0.063 ± 0.042	0.117 ± 0.074	-0.120 ± 0.178
	Int. Gradients	0.096 ± 0.030	0.174 ± 0.048	0.017 ± 0.068
	LIME	0.120 ± 0.044	0.212 ± 0.071	0.033 ± 0.111
ResNet50	Grad-CAM	0.165 ± 0.098	0.273 ± 0.139	0.209 ± 0.229
	Grad-CAM++	0.164 ± 0.091	0.272 ± 0.133	0.206 ± 0.221
	Int. Gradients	0.120 ± 0.046	0.211 ± 0.073	0.106 ± 0.140
	LIME	0.143 ± 0.056	0.247 ± 0.087	0.103 ± 0.126
MobileNetV3-Large	Grad-CAM	0.158 ± 0.007	0.272 ± 0.010	0.000 ± 0.000
	Grad-CAM++	0.158 ± 0.007	0.272 ± 0.010	0.000 ± 0.000
	Int. Gradients	0.089 ± 0.012	0.163 ± 0.020	0.007 ± 0.038
	LIME	0.129 ± 0.034	0.227 ± 0.052	0.040 ± 0.072
<i>Overall mean</i>		0.126 ± 0.059	0.219 ± 0.091	0.051 ± 0.160

Overall drift severity. All three architectures exhibit substantial explainability drift. The overall mean top- k IoU of 0.126 ± 0.059 indicates limited spatial overlap between FP32 and INT8 top-saliency masks—well below the IoU = 0.5 threshold commonly considered meaningful overlap. This finding directly challenges the common assumption that accuracy-preserving quantization also preserves the spatial structure of model explanations.

Architecture dependence. ResNet50 shows the highest mean IoU (0.148) and Spearman correlation (0.156), suggesting that its standard residual architecture preserves more coarse spatial structure during quantization, albeit with high inter-image variability (IoU std = 0.077). EfficientNetV2-S shows the tightest distribution (IoU std = 0.050) but the lowest absolute overlap (IoU = 0.096). MobileNetV3-Large shows moderate IoU (0.133) with near-zero Spearman correlation ($\rho = 0.012$), indicating that what appears to be preserved spatial overlap is in fact coincidental rather than reflecting genuine attribution correspondence.

MobileNetV3 Grad-CAM collapse. The most striking finding is the complete collapse of gradient-based attribution for MobileNetV3-Large under INT8 quantization. Both Grad-CAM and Grad-CAM++ yield *identical* drift metrics across *all* 20 evaluation images (IoU = 0.158 ± 0.007 , Spearman $\rho = 0.000$), with the only variation arising from the spatial dimensions of the binarised masks. The Spearman correlation of precisely zero—arising from constant-valued saliency maps where all pixels receive tied ranks—indicates that the fake-quantized model produces spatially uniform activation maps that destroy all class-discriminative attribution information. This collapse occurs on 20/20 evaluation images for both Grad-CAM and Grad-CAM++ (i.e., a 100% collapse rate). It is unique to MobileNetV3’s depthwise-separable convolutions, where the final convolutional layer’s channel-wise activations are independently quantized; the resulting quantised feature map degenerates into a near-constant spatial pattern. The observed IoU ≈ 0.158 reflects the top- k

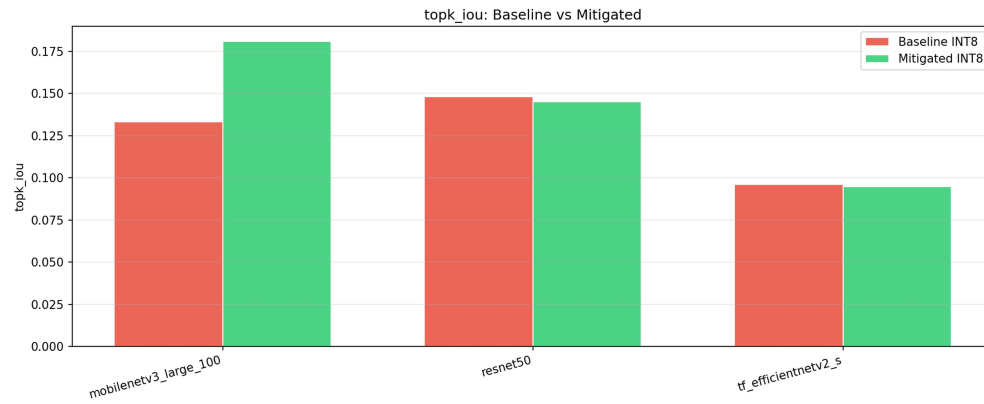
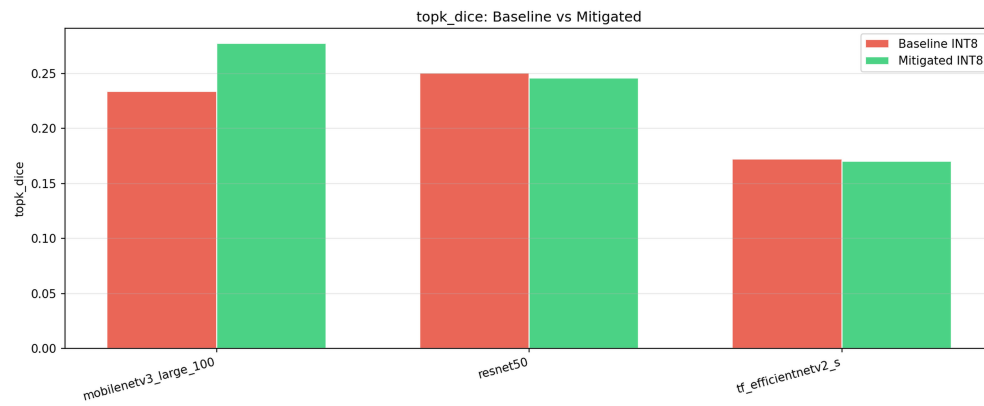
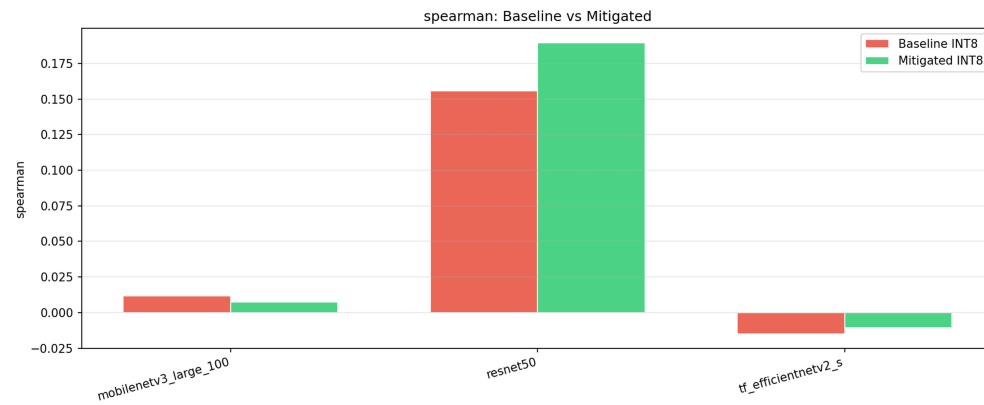
(a) Top- k IoU(b) Top- k Dice(c) Spearman ρ

Figure 6. Explainability drift comparison for (a) Top- k IoU, (b) Top- k Dice, and (c) Spearman ρ across architectures and XAI methods.

binarisation outcome when the INT8 saliency map is constant-valued (all pixels tied), not meaningful spatial agreement. Importantly, this occurs despite MobileNetV3 achieving the *highest* prediction agreement (90%), illustrating a critical disconnect between prediction stability and explanation stability.

XAI method robustness ranking. Table 7 summarises the cross-architecture robustness of each XAI method.

Table 7. XAI method robustness ranking averaged across all three architectures. CV: coefficient of variation (σ/μ); lower CV indicates more stable drift behaviour.

XAI Method	Mean IoU	Mean ρ	IoU Std	IoU CV
Grad-CAM	0.143	0.073	0.071	0.50
LIME	0.131	0.059	0.046	0.35
Grad-CAM++	0.128	0.029	0.074	0.57
Int. Gradients	0.102	0.043	0.035	0.34

LIME demonstrates the most consistent drift behaviour (CV = 0.35) and competitive Spearman correlation, suggesting that perturbation-based methods are inherently more robust to quantization than gradient-based methods. This robustness stems from LIME’s black-box query mechanism, which is unaffected by internal gradient distortions. Integrated Gradients shows the lowest IoU (0.102) but the most stable performance across images (CV = 0.34), consistent with its path-based integration providing a natural smoothing effect that dampens the impact of individual weight perturbations. Grad-CAM yields the highest overall IoU (0.143) but with the largest variance, reflecting its sensitivity to the specific activations in the last convolutional layer.

Per-class drift analysis

The 10-class structure of the Paddy Disease dataset enables fine-grained analysis of how disease-specific visual characteristics modulate drift severity. Figures 7–9 present per-image, per-method Spearman correlation heatmaps for each architecture, each cell encoding the rank-order preservation (ρ) between FP32 and fake-quantized saliency maps for a specific evaluation image and XAI method.

For EfficientNetV2-S (Figure 7), the grid reveals a heterogeneous drift landscape with pronounced method-dependent structure. Grad-CAM++ exhibits the most severe attribution degradation, with multiple images displaying negative Spearman values—indicating that INT8 quantization does not merely add noise but can systematically invert pixel attribution rankings. In contrast, LIME consistently yields the highest per-image correlations, confirming its robustness as a perturbation-based method unaffected by internal gradient distortions. At the class level, diseases with localised, high-contrast lesion patterns show the highest preservation: class 0 (bacterial leaf blight, mean IoU = 0.115) and class 5 (dead heart, IoU = 0.122). Diseases with diffuse, low-contrast symptoms exhibit the most severe drift: class 6 (downy mildew, IoU = 0.076) and class 9 (tungro, IoU = 0.091). This pattern supports the hypothesis that concentrated, high-contrast features produce sharper activation peaks more robust to quantization rounding, while spatially diffuse symptoms generate distributed activations whose subtle inter-pixel gradients are more easily disrupted.

For ResNet50 (Figure 8), the grid exhibits the highest inter-image variability among all architectures, with individual Spearman values spanning from near-zero to moderate positive correlation. Grad-CAM and Grad-CAM++ track each other closely across individual images—reflecting their shared dependence on the same final convolutional layer gradients—producing visually correlated column pairs. The per-class pattern diverges from EfficientNetV2-S: class 8 (normal/healthy leaves, IoU = 0.185) and class 5 (dead heart, IoU = 0.195) show the highest preservation, while class 7 (hispa, IoU = 0.070) exhibits the most severe drift. The relatively high preservation for healthy leaves likely reflects their uniform visual appearance, which produces broad, low-specificity saliency maps inherently less sensitive to fine-grained weight perturbations. The high variability across images suggests that ResNet50’s standard convolutions preserve some gradient diversity under quantization, but the degree of preservation is strongly image-dependent rather than architecture-guaranteed.

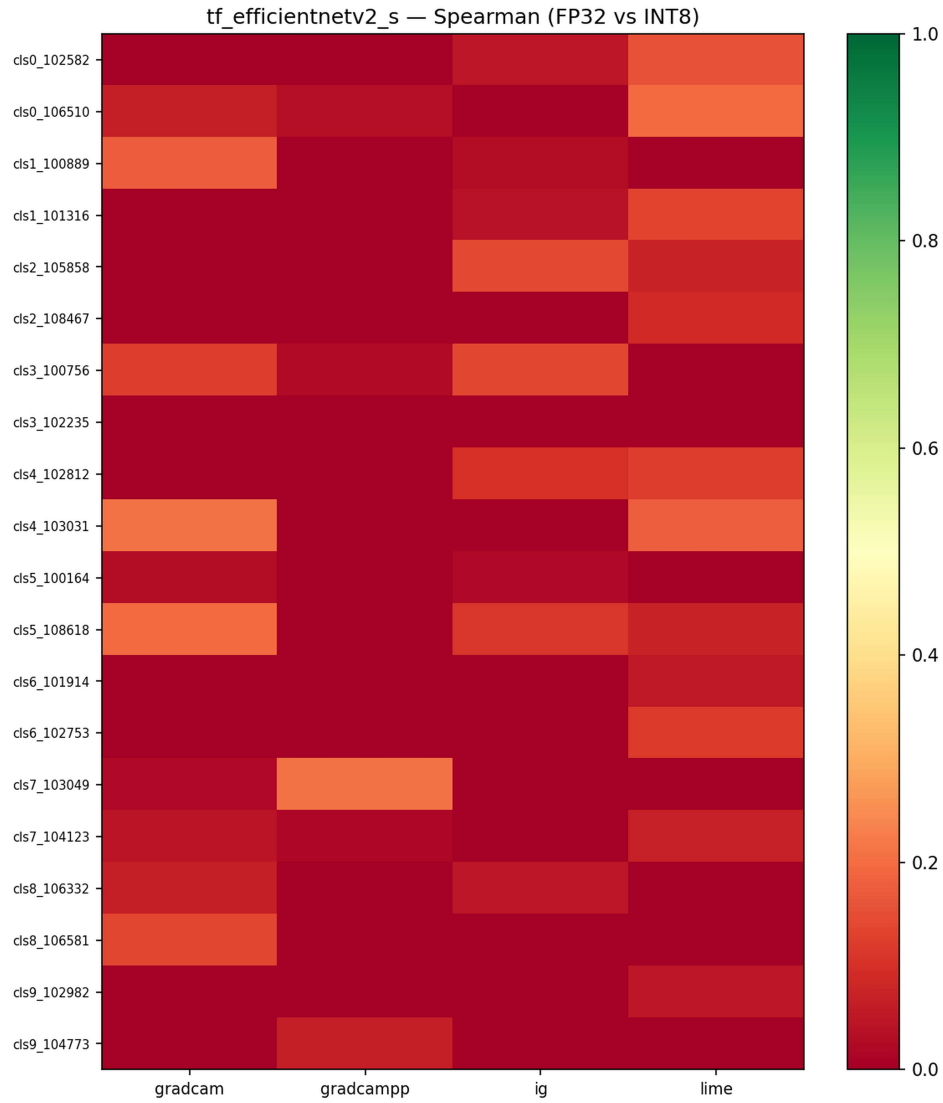


Figure 7. Per-image Spearman correlation grid for EfficientNetV2-S. Each cell shows the Spearman ρ between FP32 and fake-quantized saliency maps for a specific image–method combination. Warm colours indicate better rank preservation; cool colours indicate rank inversion.

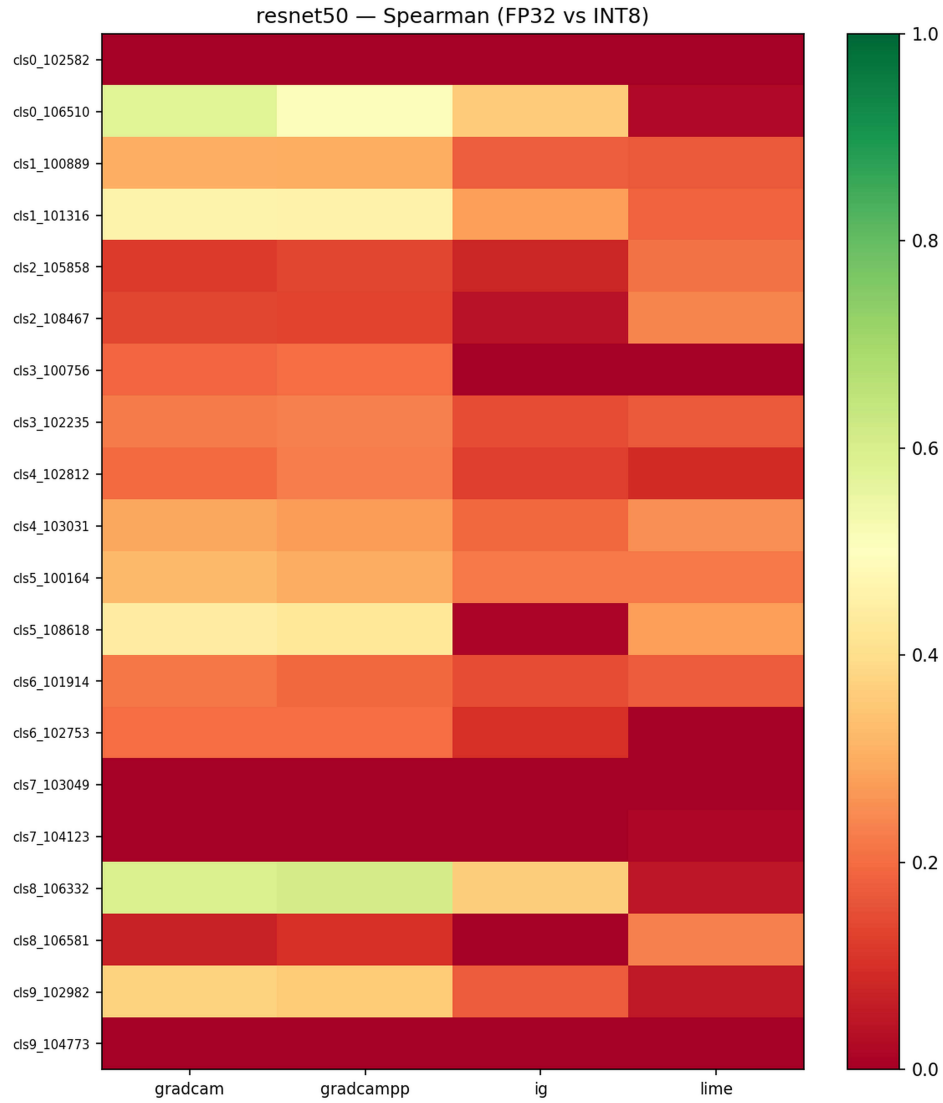


Figure 8. Per-image Spearman correlation grid for ResNet50. Notation follows Figure 7. ResNet50 shows the highest per-image variability across XAI methods.

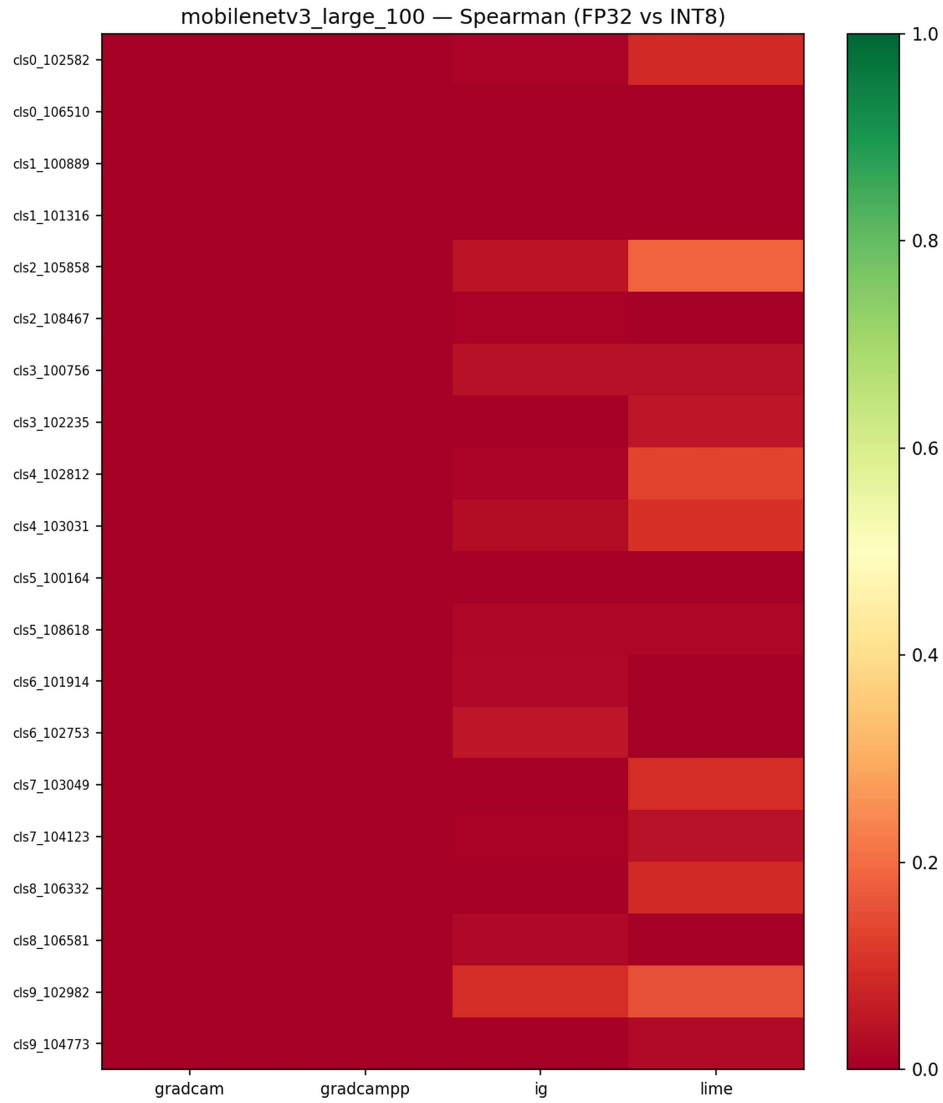


Figure 9. Per-image Spearman correlation grid for MobileNetV3-Large. Notation follows Figure 7. Grad-CAM and Grad-CAM++ columns are uniformly zero, indicating complete attribution collapse.

MobileNetV3-Large (Figure 9) produces the most visually striking grid pattern: the Grad-CAM and Grad-CAM++ columns are uniformly zero across all 20 evaluation images, forming a stark bipartite structure against the non-zero LIME and Integrated Gradients columns. This single-image-level confirmation of the Grad-CAM collapse demonstrates that the phenomenon is not an artefact of averaging but a deterministic consequence of depthwise-separable quantization. The surviving LIME column retains sparse but meaningful positive values, confirming that the model’s input–output mapping is preserved even when internal gradient pathways are destroyed. MobileNetV3-Large also shows the most uniform per-class drift distribution (IoU range: 0.124–0.142): the structural collapse of gradient-based maps eliminates class-dependent variation, and the residual variation is driven entirely by LIME and Integrated Gradients. This architecture-specific pattern has direct deployment implications—practitioners using MobileNet-family architectures on edge devices must rely exclusively on perturbation-based methods for post-quantization explainability. Supplementary Table S2 provides per-class drift metrics for all three architectures, with the caveat that only 2 images per class were evaluated (directional evidence).

QAT mitigation effectiveness

Table 8 compares baseline PTQ drift with post-QAT drift for all three architectures.

Table 8. QAT mitigation results showing mean drift values before and after XAI-preserving quantization-aware training. Δ : relative improvement (%); positive values indicate improved preservation.

Architecture	Method	Top- k IoU			Top- k Dice			Spearman ρ		
		Base	QAT	$\Delta\%$	Base	QAT	$\Delta\%$	Base	QAT	$\Delta\%$
EfficientNetV2-S	Grad-CAM	0.105	0.097	−7.8	0.185	0.173	−6.8	0.010	0.006	−40.9
	Grad-CAM++	0.063	0.072	+12.9	0.117	0.131	+12.3	−0.120	−0.061	+49.2
	Int. Grad.	0.096	0.097	+0.6	0.174	0.175	+0.4	0.017	0.005	−73.0
	LIME	0.120	0.114	−4.8	0.212	0.203	−4.3	0.033	0.008	−77.1
ResNet50	Grad-CAM	0.165	0.149	−10.1	0.273	0.248	−9.0	0.209	0.248	+18.8
	Grad-CAM++	0.164	0.149	−9.5	0.272	0.248	−8.9	0.206	0.245	+19.1
	Int. Grad.	0.120	0.124	+3.6	0.211	0.218	+3.1	0.106	0.141	+33.8
	LIME	0.143	0.160	+11.6	0.247	0.269	+9.0	0.103	0.123	+19.8
MobileNetV3	Grad-CAM	0.158	0.243	+53.9	0.272	0.346	+27.1	0.000	0.000	—
	Grad-CAM++	0.158	0.243	+53.9	0.272	0.346	+27.1	0.000	0.000	—
	Int. Grad.	0.089	0.095	+6.7	0.163	0.172	+5.7	0.007	0.009	+33.8
	LIME	0.129	0.143	+11.3	0.227	0.244	+7.7	0.040	0.021	−46.9

Selective and inconsistent improvement. QAT mitigation yields non-uniform improvements across architectures and XAI methods. For ResNet50, Spearman correlation consistently improves across all four methods (+18.8% to +33.8%), suggesting that the residual architecture’s gradient flow is amenable to CAM-consistency fine-tuning. However, this Spearman improvement comes at the cost of a modest IoU reduction for Grad-CAM (−10.1%) and Grad-CAM++ (−9.5%), indicating that QAT reranks attribution intensity ordering without necessarily expanding the spatial overlap of top salient regions.

For EfficientNetV2-S, the results are more mixed: Grad-CAM++ shows the largest improvement (+12.9% IoU, +49.2% Spearman), moving from the worst-performing to a moderately preserved method. However, standard Grad-CAM degrades (−7.8% IoU), and LIME’s Spearman correlation worsens substantially (−77.1%), suggesting that the single-method CAM teacher loss creates adversarial dynamics where optimising for Grad-CAM++ preservation disrupts other XAI methods.

For MobileNetV3-Large, QAT increases IoU by 53.9% for both Grad-CAM and Grad-CAM++, but the Spearman correlation remains at exactly 0.000. This confirms that the activation-map collapse is *structural*, arising from the architecture’s depthwise-separable design rather than from weight-value sensitivity, and cannot be resolved through training-based mitigation alone. The IoU increase reflects a shift in the constant

activation pattern from $\sim 15.8\%$ to $\sim 24.3\%$ pixel overlap with the FP32 mask, rather than any restoration of discriminative spatial attribution.

Discussion

This study establishes that INT8 post-training quantization induces substantial and systematic explainability drift across all tested architectures, XAI methods, and disease classes. The observed low top- k IoU indicates limited overlap between FP32 and INT8 top-saliency regions, challenging the assumption that accuracy-preserving quantization preserves explanations. We discuss the implications, mechanistic insights, limitations, and future directions below.

Drift is architecture-dependent and can be catastrophic

The most significant finding is the complete collapse of gradient-based attribution (Grad-CAM, Grad-CAM++) for MobileNetV3-Large under INT8 quantization, producing spatially uniform saliency maps (Spearman $\rho = 0.000$) that carry no class-discriminative information. Critically, this occurs despite MobileNetV3 achieving the *highest* prediction agreement (90%)—meaning practitioners would observe correct predictions accompanied by meaningless explanations. This disconnect between prediction stability and explanation stability is a previously unidentified hazard of quantized deployment.

The architectural origin of this collapse lies in the interaction between MobileNetV3’s depthwise-separable convolutions and per-channel INT8 quantization. In standard convolutions, each output channel integrates information across all input channels, producing spatially diverse feature maps with rich gradient structure. In depthwise-separable layers, each channel is convolved independently before pointwise projection; per-channel INT8 quantization rounds the narrow dynamic range of depthwise activations to very few distinct integer levels, collapsing the spatial diversity of the resulting feature maps. Since Grad-CAM computes class activation maps from these final-layer activations, spatially uniform feature maps yield correspondingly uniform attribution maps regardless of gradient quality.

This mechanism is consistent with the known quantization sensitivity of depthwise-separable layers[2]: because each depthwise channel is convolved independently with a narrow dynamic range, per-channel INT8 quantization collapses the spatial diversity of activations to very few distinct integer levels, producing near-constant feature maps. Since Grad-CAM computes class activation maps directly from these activations, spatially uniform feature maps yield correspondingly uniform attribution maps regardless of gradient quality. The fake-quantization framework confirms that this is an activation-level phenomenon rather than a gradient-path artefact, as gradients flow unimpeded through the straight-through estimator while activations are genuinely quantized.

This mechanism—confirmed by our use of fake-quantized models where gradients flow unimpeded through the straight-through estimator while activations are genuinely quantized—has direct implications for the mobile-deployment use case that MobileNetV3 was designed for: practitioners selecting MobileNet architectures for edge deployment must be aware that gradient-based XAI is fundamentally unreliable under INT8 quantization for this architecture family.

In contrast, ResNet50’s standard residual architecture preserves more gradient diversity (mean Spearman $\rho = 0.156$), and EfficientNetV2-S occupies an intermediate position. Notably, when restricting the analysis to the matched-prediction subset—where FP32 and quantized models agree on the predicted class—drift patterns remain qualitatively unchanged, confirming that the observed explainability drift is driven by quantization-induced changes to the attribution mechanism itself rather than by changes in the explained decision. These results demonstrate that architecture selection—typically guided solely by accuracy–efficiency trade-offs—must also consider explainability robustness when deployment requires trustworthy explanations.

Perturbation-based methods are more quantization-robust

LIME consistently outperforms gradient-based methods in drift stability (IoU CV = 0.35), which is expected given its model-agnostic, black-box query mechanism. Since LIME constructs explanations by perturbing

the input and observing output changes, it is unaffected by the quantization-induced distortions in intermediate feature maps and gradients that compromise gradient-based methods. Notably, LIME’s Spearman correlation remains positive across all architectures, including MobileNetV3-Large, confirming that the model-agnostic approach circumvents the gradient-collapse pathology entirely.

Integrated Gradients, despite being gradient-based, shows the lowest drift variance ($CV = 0.34$), likely because its path integral across 50 interpolation steps provides a natural smoothing effect that dampens the impact of individual weight perturbations. This smoothing interpretation is consistent with the theoretical analysis of Sundararajan et al.[6], who showed that path-integrated attributions satisfy completeness and sensitivity axioms that single-pass gradient methods violate.

These findings provide actionable guidance for deployment: when deploying quantized models with XAI requirements, perturbation-based methods (LIME) or path-integrated gradient methods (Integrated Gradients) should be preferred over single-pass gradient methods (Grad-CAM, Grad-CAM++).

QAT mitigation: promising but architecture-dependent

The XAI-preserving QAT with teacher–student CAM consistency loss achieves meaningful improvements for specific method–architecture combinations (e.g., +49.2% Spearman for EfficientNetV2-S Grad-CAM++, +33.8% Spearman for ResNet50 Integrated Gradients). ResNet50 exhibits the most consistent improvement, with Spearman correlation increasing for all four XAI methods (+18.8% to +33.8%). This suggests that architectures with straightforward gradient flow (residual connections without depthwise separation) are most amenable to CAM-consistency regularisation.

However, the improvements are not universal. For EfficientNetV2-S, optimising for Grad-CAM++ preservation degrades other methods (−77.1% Spearman for LIME), suggesting that single-method CAM losses create adversarial dynamics across explanation types. For MobileNetV3-Large, QAT increases IoU by 53.9% but Spearman remains at zero, confirming that the activation-map collapse is structural and cannot be resolved through fine-tuning alone.

These findings align with recent work on quantization-aware training[2] and extend it to the explainability domain. We note that the present study focuses exclusively on explainability drift metrics; formal re-benchmarking of classification accuracy (F1, balanced accuracy) after QAT was not performed, and QAT-induced accuracy changes—while expected to be small under the $\lambda = 0.5$ loss weighting—should be verified before deployment. Future mitigation strategies should explore: (i) multi-method fusion losses that simultaneously preserve multiple XAI methods; (ii) architecture-specific interventions such as retaining depthwise-separable layers in FP32 (mixed-precision strategies); and (iii) learnable quantization parameters that are explicitly optimised for gradient preservation.

Implications for trustworthy edge deployment

In precision agriculture, agronomists typically verify model predictions by inspecting saliency maps overlaid on leaf images, seeking confirmation that the model attends to disease lesion regions rather than background elements[8]. Our results demonstrate that this verification step is unreliable for quantized models: even when the model correctly identifies the disease, the explanation may point to entirely different image regions ($IoU < 0.15$) or provide no spatially discriminative information whatsoever (MobileNetV3 collapse). This finding has direct implications for regulatory frameworks that mandate explainability for high-risk AI systems and suggests that post-quantization explainability auditing should be incorporated into model deployment checklists alongside standard accuracy validation.

The per-class analysis further reveals that clinically important distinctions in drift severity may go undetected in aggregate metrics. For EfficientNetV2-S, the $1.6\times$ difference in IoU between the best-preserved class (dead heart, $IoU = 0.122$) and worst-preserved class (downy mildew, $IoU = 0.076$) means that quantization disproportionately compromises explanations for diseases with subtle, diffuse symptoms—precisely the cases where expert verification through XAI is most valuable.

Connection to prior work

Our findings complement and extend several strands of prior research. Hooker et al.[48] demonstrated that pruning selectively degrades features important for underrepresented classes; we observe an analogous

phenomenon with quantization, where classes with diffuse visual patterns suffer greater drift. The ECQx framework[28] and related work[29, 34] used XAI to *guide* quantization decisions; our results provide the complementary perspective of how standard quantization *disrupts* XAI outputs. The MobileNetV3 Grad-CAM collapse we document is, to our knowledge, the first reported instance of complete gradient-based attribution failure under quantization, extending Nagel et al.’s[2] observation that depthwise-separable layers are “quantization-sensitive” to the explainability domain.

Limitations

Several limitations should be acknowledged. First, our evaluation uses 20 validation images (two per class), which, while sufficient to identify major drift patterns and the MobileNetV3 collapse, may not capture the full distribution of drift behaviour; larger-scale evaluation would strengthen statistical power. Second, drift metrics were computed using fixed top- k binarisation thresholds, which may mask drift in lower-activation regions that are still clinically meaningful. Third, the study evaluates only INT8 PTQ; other quantization bit-widths (INT4, FP16) and schemes (mixed-precision, FP8) may exhibit different drift profiles. Fourth, ground-truth disease masks were not available for the Paddy Disease dataset, preventing assessment of whether FP32 or INT8 saliency maps better align with actual disease regions—a question that would determine whether drift is always deleterious or, in some cases, serendipitously beneficial. Fifth, although the PyTorch fake-quantization framework closely simulates INT8 arithmetic (and we verified prediction equivalence with the ONNX Runtime INT8 model), minor numerical differences between simulated and true integer arithmetic may exist.

Future work

We identify four primary directions. First, extending the drift benchmark to additional architectures (Vision Transformers, ConvNeXt), quantization schemes (INT4, FP8, mixed-precision), and domains (medical imaging, autonomous driving, remote sensing). Second, developing architecture-aware mitigation strategies that account for layer-specific quantization sensitivity—for example, preserving depthwise-separable layers in FP32 for MobileNet-family architectures while quantizing remaining layers to INT8. Third, investigating the relationship between explainability drift and faithfulness[48]: if drift moves saliency maps closer to ground-truth pathology regions (e.g., by suppressing noisy background attributions), traditional notions of drift as purely negative may require revision. Fourth, designing multi-method XAI consistency losses for QAT that simultaneously preserve the spatial attribution patterns of gradient-based, path-integrated, and perturbation-based explanation methods.

Conclusion

This study establishes *explainability drift* as a critical, previously unquantified consequence of INT8 post-training quantization in deep learning models for plant disease detection. Through systematic evaluation of four post-hoc XAI methods (Grad-CAM, Grad-CAM++, Integrated Gradients, LIME) across three architectures (EfficientNetV2-S, ResNet50, MobileNetV3-Large) on the Paddy Disease Classification dataset, we demonstrate that quantization-induced accuracy preservation does not guarantee explanation preservation. All architectures exhibit mean top- k IoU below 0.17, indicating low spatial overlap between FP32 and INT8 top-saliency regions after PTQ.

We identify a complete gradient-based attribution collapse in MobileNetV3-Large (Spearman $\rho = 0.000$), a novel architectural vulnerability arising from depthwise-separable convolutions under per-channel quantization. LIME and Integrated Gradients exhibit greater robustness to quantization than single-pass gradient methods, providing actionable guidance for practitioners deploying quantized models with explainability requirements. XAI-preserving quantization-aware training with a teacher–student CAM consistency loss yields selective improvements—up to 49.2% Spearman gain for specific method–architecture combinations—but remains architecture-dependent and unable to resolve the structural collapse observed in MobileNetV3.

These findings have direct implications for deploying trustworthy AI on resource-constrained edge devices in precision agriculture and other safety-critical domains. We recommend incorporating post-quantization explainability audits into the model deployment pipeline, alongside accuracy validation, to ensure that quantized models provide reliable and meaningful explanations to end users.

Data availability

The Paddy Disease Classification dataset is publicly available via Kaggle, subject to Kaggle's competition access terms and Terms of Use[46]. All code for model training, quantization, XAI computation, and drift metric evaluation will be released in a public GitHub repository upon acceptance of this manuscript.

Code availability

All analysis code, including training scripts, quantization pipelines, XAI implementations, and drift metric computation, will be released in a public GitHub repository under the MIT Licence upon acceptance.

Acknowledgements

The author thanks the Kaggle “Paddy Doctor: Paddy Disease Classification” organizers and data providers for making the dataset publicly available. Computational resources were provided by Google Colab (T4 GPU).

Author contributions statement

S.K. conceived the study, designed and conducted the experiments, implemented all training, quantization, and XAI pipelines, performed the statistical analysis, and wrote the manuscript.

Funding

Not applicable.

Competing interests

The author declares no competing interests.

References

- [1] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630*, 2021. doi: 10.48550/arXiv.2103.13630.
- [2] Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021. doi: 10.48550/arXiv.2106.08295.
- [3] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2704–2713, 2018. doi: 10.1109/CVPR.2018.00286.
- [4] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.
- [5] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. doi: 10.1109/WACV.2018.00097.

- [6] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3319–3328, 2017.
- [7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why should I trust you?”: explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016. doi: 10.1145/2939672.2939778.
- [8] Koushik Nagasubramanian, Sarah Jones, Asheesh K. Singh, Soumik Sarkar, Arti Singh, and Baskar Ganapathysubramanian. Plant disease identification using explainable 3D deep learning on hyperspectral images. *Plant Methods*, 15:98, 2019. doi: 10.1186/s13007-019-0479-8.
- [9] Mingxing Tan and Quoc V. Le. EfficientNetV2: smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 10096–10106, 2021.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.
- [11] Andrew Howard, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, Ruoming Pang, Hartwig Adam, and Quoc Le. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, 2019. doi: 10.1109/ICCV.2019.00140.
- [12] Ting Qin, Zhao Li, Jiaqi Zhao, Yuting Yan, and Yafei Du. Mixed precision quantization based on information entropy. *Scientific Reports*, 15:91684, 2025. doi: 10.1038/s41598-025-91684-8.
- [13] Archisman Biswas, Raghuvar Singh, Suresh Elangovan, Simran Sabnis, and Udayan Ganguly. Regularization-based framework for quantization-, fault- and variability-aware training. *arXiv preprint arXiv:2502.03971*, 2025. doi: 10.48550/arXiv.2502.03971.
- [14] Zheyuan Lu et al. Efficient quantization-aware training on Segment Anything Model for medical images. *arXiv preprint*, 2024.
- [15] Minjun Kim, Jiwon Choi, Jungwoo Lee, Wonhyuk Cho, and U Kang. Zero-shot quantization: a comprehensive survey. In *Proceedings of the IJCAI-25 Survey Track*, 2025.
- [16] Dendi Maulana and Rajasvaran Kamalasvaran Ramasamy. Systematic review of quantization-optimized lightweight Transformer architectures for real-time fruit ripeness detection on edge devices. *Computers*, 15(1):69, 2026. doi: 10.3390/computers15010069.
- [17] Nusrat Nigar, Mohammad Imran Faisal, Muhammad Umer, Oluwafemi Oki, and Joel Lukose. Improving plant disease classification with deep learning based prediction model using explainable artificial intelligence. *IEEE Access*, 12:99741–99754, 2024. doi: 10.1109/ACCESS.2024.3428553.
- [18] A. Porfirio, Rui Neves Madeira, and Cláudia Santos. Exploring explainable AI techniques for plant disease classification in digital agriculture. *Procedia Computer Science*, 265:175–182, 2025. doi: 10.1016/j.procs.2025.07.170.
- [19] Md. Osman Gani et al. PapayaNet: an attention-guided CNN for papaya disease classification with Grad-CAM and LIME explanations. *Scientific Reports*, 15:25374, 2025. doi: 10.1038/s41598-025-25374-w.
- [20] Md. Nurullah et al. Multi-label tomato disease detection using Vision Transformer with explainable AI. *Electronics*, 14(23):4762, 2025. doi: 10.3390/electronics14234762.
- [21] Ayesha Siddika Preotee et al. Bangladeshi leaf disease detection using explainable deep learning. In *Proceedings of the 27th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2024. doi: 10.1109/ICCIT64611.2024.11022335.

- [22] Abrar Nawar et al. EfficientNet-B0 based plant disease classification using SHAP explainability. In *Proceedings of the International Conference on Computing Advancements (ICCA)*. ACM, 2024. doi: 10.1145/3723178.3723315.
- [23] Md. Rakibul Hasan, Md. Enamul Haque, Nitun Sarker, Tawsin Arafat Abdullah, Sourav Roy, Md. Asaduzzaman Inam Taaha, Syed Abu Taher Arafat, Md. Jaber Al Patwary, Mominul Ahsan, and Julfikar Haider. Mulberry leaf disease detection by CNN-ViT with XAI integration. *PLoS ONE*, 20(4):e0325188, 2025. doi: 10.1371/journal.pone.0325188.
- [24] Pranav Thakur et al. PlantXViT: a hybrid Vision Transformer–CNN for explainable plant disease classification. *arXiv preprint*, 2022.
- [25] Faria Tabassum and Boshir S. Ahamed. CBAM-enhanced betel leaf disease classification with Grad-CAM explainability. In *Proceedings of the International Conference on Big Data, IoT and Machine Learning (BIM)*. Springer, 2024.
- [26] Nikita A. Lysov, Ekaterina I. Maximova, Evgeny P. Vasiliev, Anna A. Getmanskaya, and Vadim E. Turlapov. Entropy as a high-level feature for XAI-based early plant stress detection. *Entropy*, 24(11): 1597, 2022. doi: 10.3390/e24111597.
- [27] R. Kusuma and R. Rajkumar. Plant leaf disease detection and classification using artificial intelligence techniques: a review. *Indonesian Journal of Electrical Engineering and Computer Science*, 38(2): 1308–1323, 2025. doi: 10.11591/ijeecs.v38.i2.pp1308-1323.
- [28] Daniel Becking, Maximilian Dreyer, Wojciech Samek, Klaus-Robert Müller, and Bjarne Barz. ECQx: explainability-driven quantization for low-bit and sparse DNNs. In *Explainable Artificial Intelligence*, volume 13200 of *Lecture Notes in Artificial Intelligence*, pages 271–296. Springer, 2022. doi: 10.1007/978-3-031-04083-2_14.
- [29] Rakesh Ranjan and Andreas Savakis. LRP-QViT: mixed-precision Vision Transformer quantization via layer-wise relevance propagation. *arXiv preprint arXiv:2401.11243*, 2024. doi: 10.48550/arXiv.2401.11243.
- [30] Rakesh Ranjan and Andreas Savakis. Mix-QViT: mixed-precision Vision Transformer quantization driven by layer importance and quantization sensitivity. *arXiv preprint*, 2025.
- [31] Ryo Tashiro and Hiromitsu Awano. Pay attention via quantization: enhancing explainability of neural networks via quantized activation. *IEEE Access*, 11:34437–34447, 2023. doi: 10.1109/ACCESS.2023.3264855.
- [32] Amin Maleki, Masoud Lavaei, Mohammadhashem Bagheritabar, Shervin Beigzad, and Zahra Abadi. Quantized and interpretable learning scheme for deep neural networks in classification task. *arXiv preprint arXiv:2412.05319*, 2024. doi: 10.48550/arXiv.2412.05319.
- [33] Elham Manouchehri Rezabeyk, Shervin Beigzad, Yasaman Hamzavi, Mohammadhashem Bagheritabar, and Seyed Salar Mirikhoozani. Saliency assisted quantization for neural networks. *arXiv preprint arXiv:2411.05850*, 2024. doi: 10.48550/arXiv.2411.05850.
- [34] Minjun Kim, Jiyun Kim, and U Kang. SYNQ: accurate zero-shot quantization by synthesis-aware fine-tuning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.
- [35] Han Xiao, Yi-Lin Sung, Elias Stengel-Eskin, and Mohit Bansal. Task-circuit quantization: leveraging knowledge localization and interpretability for compression. *arXiv preprint arXiv:2504.03970*, 2025. doi: 10.48550/arXiv.2504.03970.
- [36] Kazuki Otsuki, Junpei Iida, Noe Doublet, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi, and Hiroshi Sugiura. Layer-wise relevance propagation with conservation property for ResNet. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2024.

- [37] Rohit Malik, Dhruv Seth, Paarth Singh, Yash Chitroda, and Harish K. Sankarapu. Interpretability-aware pruning for efficient medical image analysis. *arXiv preprint*, 2025.
- [38] Md. Johirul Hasan et al. 3D MRI brain tumor classification using explainable deep learning with channel attention. *Expert Systems with Applications*, page 131513, 2026. doi: 10.1016/j.eswa.2026.131513.
- [39] Xuan Wang, Jing Guan, Shuang Li, et al. Explainable-enhanced AI for diagnosing coronary microvascular dysfunction with multimodal imaging. *iScience*, 28:114101, 2025. doi: 10.1016/j.isci.2025.114101.
- [40] Cynthia L. Bennett, Heather Cole-Lewis, et al. Interdisciplinary expertise to advance human-centered explainable AI. *Social Science & Medicine*, 394:119036, 2026. doi: 10.1016/j.socscimed.2026.119036.
- [41] Antonio De Fazio et al. CARE: a rule-extraction framework for PUV diagnosis. *Journal of Systems and Software*, page 112720, 2026. doi: 10.1016/j.jss.2025.112720.
- [42] David McDonnell et al. Explainable AI for insurance claims prediction. *Technological Forecasting and Social Change*, page 124418, 2026. doi: 10.1016/j.techfore.2025.124418.
- [43] Hiroyuki Yoshida et al. Estimating Fukushima dose rates using LightGBM with SHAP explainability. *Journal of Environmental Radioactivity*, page 107900, 2026. doi: 10.1016/j.jenvrad.2026.107900.
- [44] Sergio Bañales, Raúl Dormido, and Natividad Duro. Explainable AI for predicting household demand flexibility. *Energy and AI*, 24:100686, 2026. doi: 10.1016/j.egyai.2026.100686.
- [45] Shunlei Guo, Yi Zhong, Zhiyong Feng, Song Kang, and Jiajie Chen. Towards transparent 6G AI-RAN: a survey on explainable deep reinforcement learning for intelligent network slicing. *Journal of Information and Intelligence*, 2025. doi: 10.1016/j.jiixd.2025.12.005.
- [46] Loke Gokul and P. Anantharaman. Paddy Doctor: Paddy disease classification. Kaggle Competition Dataset, 2022. URL <https://www.kaggle.com/competitions/paddy-disease-classification>. Accessed: 2024.
- [47] Ross Wightman. PyTorch Image Models. <https://github.com/rwightman/pytorch-image-models>, 2021.
- [48] Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2020. doi: 10.48550/arXiv.1911.05248.