International Journal of Advanced Engineering and Management Research

Vol. 11, No. 04; 2026

ISSN: 2456-3676

Interpretable Image-level Acne Severity Grading via Efficientnet-B0 Transfer Learning and Grad-CAM: An Open, Cross-platform Reference Implementation Validated on the ACNE04 Benchmark

Sophie Zeng[1], Sean Kalaycioglu[1,2],*, Collin Hong[1,3], Haipeng Xie[1]

[1]Dr. Robot Inc., Toronto, ON, Canada

[2]Department of Aerospace Engineering, Toronto Metropolitan University, Toronto, ON, Canada

[3]Skinopathy Inc., Toronto, ON, Canada

*Corresponding author:

doi.org/10.51505/ijaemr.2026.11406 URL: http://dx.doi.org/10.51505/ijaemr.2026.11406

Received: July 27, 2026 Accepted: Aug 03, 2026 Online Published: Aug 10, 2026

Abstract

Acne vulgaris affects most adolescents and a substantial proportion of adults; accurate severity grading determines treatment choice, monitoring, and trial endpoints, yet manual assessment on scales such as the Investigator's Global Assessment or Hayashi criteria is limited by inter-rater variability and imaging conditions. We developed and evaluated an image-level four-class Hayashi-criterion acne severity classifier based on transfer learning of Image Net-pretrained EfficientNet-B0. The model was fine-tuned on the public ACNE04 benchmark (2,983 labeled images) with AdamW optimization, standard geometric and photometric augmentation, and checkpoint selection by validation macro-F1. On a held-out stratified 15 % test set the classifier achieved 93.5 % accuracy and 94.4 % macro-F1 (per-class F1 0.92–0.97), with 83 % of errors confined to adjacent grades; Cohen's quadratic-weighted κ reached 0.956 (95 % CI [0.935, 0.973]). Bootstrap confidence intervals confirmed stable performance. Grad-CAM visualizations generated from the final convolutional block concentrated on clinically relevant facial regions (forehead, cheeks, chin). The complete pipeline is released as functionally equivalent open-source implementations in Python (PyTorch + timm) and MATLAB R2026a, including a clinician-facing inference widget and a tiered backbone fallback that guarantees runnability without specialized pretrained-weight packages. The results demonstrate that lightweight transfer learning can deliver strong, balanced, and interpretable grading performance on a public benchmark while providing a reproducible cross-platform reference for subsequent prospective and device-stratified validation required for clinical translation.

Keywords: acne severity grading, transfer learning, Efficient Net, convolutional neural networks, Grad-CAM, explainable AI, dermatology AI, computer-aided diagnosis, medical image classification, reproducible research

Vol. 11, No. 04; 2026

ISSN: 2456-3676

1. Introduction

1.1 Introduce the Problem

Acne vulgaris is among the most prevalent chronic dermatological conditions, with estimated lifetime prevalence above 80% in adolescents and meaningful persistence into adulthood (Hayashi et al., 2008; Reynolds et al., 2024). Beyond cutaneous symptoms, untreated or under-treated disease produces psychosocial morbidity, reduced quality of life, and permanent scarring, all of which scale with severity (Hayashi et al., 2008; Reynolds et al., 2024; U.S. Food and Drug Administration, 2018). Severity assessment therefore governs three downstream decisions of direct clinical consequence: choice of therapeutic class (topical vs. systemic vs. isotretinoin), monitoring of treatment response, and inclusion or endpoint definition in clinical trials. The U.S. Food and Drug Administration explicitly references the IGA as a clinical-trial endpoint structure, confirming the regulatory relevance of stable, reproducible grading (U.S. Food and Drug Administration, 2018).

In conventional practice, severity is assigned by visual inspection against one of several ordinal scales — IGA (0–4), GEA (0–5, often pooled as 0/1/2/3+/4+), Hayashi-criterion four-level grading (mild/moderate/severe/very severe), or region-specific guidelines (Bae et al., 2024; Hayashi et al., 2008; Reynolds et al., 2024). None has been universally adopted, and meta-analyses report meaningful intra- and inter-observer variation even among trained dermatologists, particularly at grade boundaries and on darker phototypes where post-inflammatory hyperpigmentation (PIH) complicates lesion identification (Bae et al., 2024; Wu et al., 2019). Lighting, camera color rendition, capture distance, facial pose, and even smartphone operating system have been shown to bias both human and machine grading (Gao et al., 2025; Huynh et al., 2022; Watanabe et al., 2025; Wen et al., 2022). The convergence of these factors — clinical importance, scale heterogeneity, and capture-condition sensitivity — has motivated a now-substantial body of work on automated acne assessment.

1.2 Limitations of Prior Automated Approaches

Existing computer-vision approaches cluster into four families. (i) *Classical image processing* — edge detection, color thresholding, contour analysis, texture features — is fast and interpretable but brittle under illumination and skin-tone variation (Huynh et al., 2022; Wen et al., 2022). (ii) *Lesion detection* with Faster R-CNN or YOLO-family detectors provides bounding-box localization that clinicians can visually verify, but requires dense per-lesion annotation (Gao et al., 2025; Huynh et al., 2022; Redmon & Farhadi, 2018; Ren et al., 2017; Wen et al., 2022; Wu et al., 2019). (iii) *Pixel-wise segmentation* (e.g., U-Net variants, often with EfficientNet encoders) delivers the finest-grained localization but demands pixel-level masks, particularly costly in dermatology where expert review is required (Gao et al., 2025; Ronneberger et al., 2015). (iv) *End-to-end image-level classification* — typically transfer learning on a CNN backbone — requires only image-level severity labels and is the most label-efficient family, at the cost of reduced transparency unless paired with an explicit explainability mechanism (Selvaraju et al., 2017; Watanabe et al., 2025).

Vol. 11, No. 04; 2026

ISSN: 2456-3676

A central methodological tension follows: detection and segmentation pipelines offer stronger interpretability at the cost of annotation burden and multi-stage error propagation, while classification pipelines are feasible with the labels most commonly available but are opaque without auxiliary explanation. Recent best practice has therefore been to pair classification with class-activation explainability — typically Grad-CAM (Selvaraju et al., 2017) — to recover a clinically inspectable signal without imposing lesion-level annotation costs.

1.3 Reproducibility, Explainability, and Deployment Gaps

A parallel literature critique is methodological rather than algorithmic. Recent reviews and primary studies converge on five reproducibility gaps that limit clinical translation: (1) inconsistent imaging-metadata reporting (device model, lens, distance, illuminance, white balance); (2) heterogeneous severity scales that complicate cross-study comparison; (3) absence of formal cross-device generalization testing despite documented iOS-vs-Android color-rendition effects on dermatologist agreement; (4) sparse adoption of clinically meaningful ordinal-agreement metrics (weighted κ, ICC) in favor of raw accuracy; and (5) limited public availability of dataset and code, preventing independent replication and external validation (Gao et al., 2025; Huynh et al., 2022; U.S. Food and Drug Administration, 2021; Watanabe et al., 2025; Wu et al., 2019). These gaps frame deployment readiness as currently bottlenecked less by raw model accuracy than by reproducibility infrastructure.

1.4 Contributions

Against this backdrop, this work makes four contributions.

- **A reproducible image-level acne severity classifier** based on EfficientNet-B0 transfer learning, trained and evaluated on the public ACNE04 benchmark (Wu et al., 2019), achieving 93.5% test accuracy and 94.4% macro-F1 with balanced per-class performance (F1 $\geq$ 0.92 for all four grades) and 95% bootstrap confidence intervals reported alongside point estimates.
- **A clinically scoped explainability component** using Grad-CAM (Selvaraju et al., 2017) applied to the final convolutional block of EfficientNet-B0, with explicit specification of the reduction layer for stable cross-backbone behavior, and a delineated interpretation envelope (heatmap localization is treated as an indirect plausibility check, not clinical proof).
- **An open, cross-platform reference implementation** in Python (PyTorch + timm) and MATLAB R2026a (Deep Learning Toolbox *imagePretrainedNetwork / trainnet / gradCAM)*, with a clinician-facing inference widget. A tiered backbone fallback (EfficientNet-B0 → ResNet-18 → MobileNet-v2 → compact from-scratch CNN) preserves end-to-end runnability without pretrained-weight support packages, lowering the entry cost for resource-constrained reviewers.
- **Transparent methodological caveats** — ordinal grade ambiguity at boundaries, single-institution data, absence of formal cross-device evaluation — explicitly framed against the regulatory and reporting expectations for AI-enabled Software as a Medical Device (SaMD)

Vol. 11, No. 04; 2026

ISSN: 2456-3676

and Good Machine Learning Practice (GMLP) (International Medical Device Regulators Forum, 2013; U.S. Food and Drug Administration, 2018, 2021).

The remainder of the paper is organized as follows. Section 2 reviews related work on acne severity scales, computer-aided acne assessment, transfer learning, and explainability. Section 3 details the dataset, model, training protocol, evaluation metrics, and cross-platform implementations. Section 4 presents quantitative and qualitative results including bootstrap confidence intervals, an ablation, and inference-cost characterization. Section 5 discusses comparison with prior work, strengths, limitations, and regulatory implications. Section 6 concludes and outlines future work.

2. Related Work

2.1 Clinical Severity Scales

Three families of acne severity scales dominate the literature (Bae et al., 2024; Hayashi et al., 2008; Reynolds et al., 2024):

- **IGA (Investigator's Global Assessment):** a 0–4 ordinal scale (clear → very severe) widely used in U.S. clinical trials (U.S. Food and Drug Administration, 2018). Some studies collapse to 0–1 / 2 / 3–4 to stabilize training under class imbalance.
- **GEA (Global Evaluation Acne):** a 0–5 European scale; severe grades (4+) are commonly pooled due to rarity in non-clinical samples.
- **Hayashi-criterion four-level grading:** mild / moderate / severe / very severe, tied to lesion-count intervals (Hayashi et al., 2008). This is the scheme used to label ACNE04 (Wu et al., 2019) and consequently the scheme adopted in this work.

A 2024 comprehensive review noted that no single scale has been universally adopted in routine clinical practice, making *ground truth* itself scale-dependent and subject to inter-rater drift (Bae et al., 2024). This drives two methodological choices in modern work: treating severity as an *ordinal* prediction problem (so that mis-grading by one level costs less than by three), and reporting *agreement metrics* (weighted κ (Landis & Koch, 1977), intraclass correlation) alongside raw accuracy.

2.2 Computer-Aided Acne Assessment

Early work used handcrafted color-thresholding and texture features (Wen et al., 2022), which are fast and interpretable but brittle to lighting and skin tone. Modern work clusters into three architectural patterns.

2.2.1 End-to-End Ordinal Grading

Standardized-imaging studies — VISIA-style rigs, fixed lighting, multi-view capture — combined with EfficientNet- or Inception-class backbones achieve high reported performance on private datasets (Watanabe et al., 2025). Watanabe et al. (2025) reported approximately 90%

Vol. 11, No. 04; 2026

ISSN: 2456-3676

accuracy and 0.885 macro-F1 on a standardized Japanese dataset using EfficientNet-B2. The strong dependence of these numbers on standardized capture is a recurring theme.

2.2.2 Two-Stage Detection Plus Grading

Pipelines such as AcneDet train Faster R-CNN (Ren et al., 2017) to detect lesions by type and then a downstream model (e.g., LightGBM) to map counts to an IGA grade (Huynh et al., 2022). Reported lesion-detection mAP is typically in the 0.4–0.6 range, with explicit class-imbalance issues for rare nodules and cysts.

2.2.3 Multi-Stage Segmentation, Lesion Scoring, and Severity Aggregation

Recent systems segment all lesions with U-Net variants (Ronneberger et al., 2015), separate touching lesions, crop lesion-centered patches, score each, and aggregate into a continuous 0–100 severity score (Gao et al., 2025). This more closely resembles clinical reasoning at the cost of substantial annotation effort.

2.2.4 Label-Distribution Learning on ACNE04

Wu et al. (2019) formulated grading and counting as a joint label-distribution-learning problem, using medical mapping intervals to link continuous lesion counts to ordinal grades. Their formulation remains the de facto benchmark on this dataset and motivates the explicitly ordinal evaluation adopted in Section 4.5.

2.3 Transfer Learning and EfficientNet

Transfer learning is near-universal in medical image classification because medical datasets are typically too small to train large networks from scratch (He et al., 2016; Tan & Le, 2019; Watanabe et al., 2025). EfficientNet is particularly attractive because its compound-scaling formulation jointly tunes depth, width, and resolution, allowing strong accuracy at modest compute (Tan & Le, 2019). EfficientNet-B0 — the smallest member of the family with approximately 5.3 M parameters — is widely used as a practical backbone for medical-imaging classification on standard-resolution inputs (224×224) and was used in this work for its favorable accuracy-to-compute ratio.

2.4 Explainability for Medical Image AI

Visual explanation methods are now considered table-stakes for medical-image AI. Grad-CAM (Selvaraju et al., 2017) computes a coarse class-discriminative localization map by combining the gradient of the predicted class score with respect to the final convolutional feature maps. The output is a single-channel attention heatmap that, when overlaid on the input image, visualizes the spatial regions most responsible for the prediction. Two caveats apply in clinical settings: (i) high attention does not imply *correct* attention — a model may attend to an irrelevant correlate that happens to discriminate classes in the training distribution; and (ii) Grad-CAM is a coarse map at the resolution of the final feature layer (typically 7×7 for ImageNet-style 224×224

Vol. 11, No. 04; 2026

ISSN: 2456-3676

inputs), not a lesion-level segmentation. Treating it as plausibility evidence rather than clinical proof is the responsible framing, and we adopt it throughout.

2.5 Cross-Device Generalization and Reproducibility

A large smartphone-based GEA study reported that inter-dermatologist agreement varied with device operating system (iOS vs. Android) even when camera specifications were nominally comparable, attributing the difference to color-rendition pipelines (Huynh et al., 2022). The implication is that *external validation must be device-stratified* to be clinically meaningful, yet this practice remains rare in published acne-AI work. Standardized clinical photography (DSLR with fixed inter-pupil distance normalization) and VISIA-style rigs sidestep the issue at the cost of accessibility. Smartphone pipelines remain attractive for triage and self-monitoring use cases but require explicit cross-device evaluation prior to deployment (Gao et al., 2025; Huynh et al., 2022).

2.6 Research Gap

The literature surveyed above leaves three concrete gaps that motivate the design choices in this work:

- **Reproducibility infrastructure.** Few published methods release end-to-end runnable code with public-dataset pointers and a clinician-facing inference interface; even fewer release the same pipeline in more than one deep-learning framework.
- **Backbone-robust explainability.** Grad-CAM implementations commonly hard-code a layer name from a single backbone, which silently fails or warns when the backbone is swapped.
- **Honest limitation framing.** Many studies report headline accuracy without discussing what cannot be inferred from internal-split evaluation (cross-device, cross-population, prospective performance) and without explicit alignment to SaMD lifecycle expectations.

The present work addresses (1) by releasing dual Python + MATLAB implementations, (2) by parameterizing the Grad-CAM reduction layer through the network's own output-node API, and (3) by an explicit limitations section aligned with the SaMD lifecycle expectations articulated by global regulators (International Medical Device Regulators Forum, 2013; U.S. Food and Drug Administration, 2018, 2021).

3. Materials and Methods

3.1 Dataset

We use the publicly released **ACNE04** dataset (Wu et al., 2019), the de facto benchmark for image-level acne grading on the Hayashi four-level scale. ACNE04 was constructed under controlled imaging conditions and provides face-level images labeled by trained dermatologists using lesion-count intervals to assign each image to one of four severity grades: Grade 0 (mild / clear), Grade 1 (moderate), Grade 2 (severe), Grade 3 (very severe). After deduplication and inclusion of the auxiliary 1024-resolution subsets distributed with the benchmark, the dataset

Vol. 11, No. 04; 2026

ISSN: 2456-3676

used in this study comprises **2,983 labeled images** distributed as 1,024 / 1,296 / 402 / 261 across Grades 0–3 (Table 1). The class distribution is imbalanced, with Grades 2 and 3 collectively representing 22% of the corpus — a property typical of population-sampled acne datasets and a known source of difficulty for both human and machine graders (Watanabe et al., 2025; Wu et al., 2019). We additionally verified the pipeline on the canonical 1,457-image ACNE04 release (497 / 637 / 186 / 137 across Grades 0–3) using the MATLAB implementation; results are consistent with those reported below within the expected variance of a smaller test set.

Table 1. Class distribution of the working dataset.

Grade	Description	Count	Percent
0	Mild / clear	1,024	34.3%
1	Moderate	1,296	43.4%
2	Severe	402	13.5%
3	Very severe	261	8.8%
Total		**2,983**	**100%**

Note. Percentages are of the 2,983-image working corpus and may not sum to 100% because of rounding.

The dataset was split stratified-by-grade into 70% training (2,088), 15% validation (447), and 15% test (448), with a fixed random seed (42) for reproducibility.

3.2 Pre-processing and Augmentation

All images were converted to RGB and resized to 224×224 to match the EfficientNet-B0 input resolution. Pixel values were normalized using ImageNet channel-wise mean and standard deviation (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) to preserve the input distribution expected by the pretrained backbone (Deng et al., 2009; Tan & Le, 2019).

Training-time augmentation comprised: random horizontal flip (p = 0.5); random rotation in ±10°; and ColorJitter with brightness 0.15, contrast 0.15, saturation 0.10. These transformations target the dominant nuisance variables identified in the cross-device literature (Huynh et al., 2022; Watanabe et al., 2025) without introducing color shifts large enough to corrupt the inflammatory-redness signal central to severity grading. Validation and test images were resized and normalized only — no random augmentation was applied, ensuring evaluation reproducibility.

3.3 Network Architecture

We adopt **EfficientNet-B0** (Tan & Le, 2019) as the convolutional backbone, initialized from ImageNet pretrained weights. The architecture comprises an initial 3×3 convolution, seven inverted-residual MBConv stages with squeeze-and-excitation modules, and a final 1×1 convolutional projection (the "head"), followed by global average pooling and a fully-connected classifier. The original 1,000-way classifier was replaced with a four-way linear layer matching

Vol. 11, No. 04; 2026

ISSN: 2456-3676

the number of acne severity grades. With ImageNet initialization the network has approximately 5.3 M parameters; only the classifier head is randomly initialized, while all earlier weights are inherited from the pretrained backbone and fine-tuned end-to-end. Figure 1 summarizes the overall pipeline architecture from input image to predicted grade and Grad-CAM overlay.

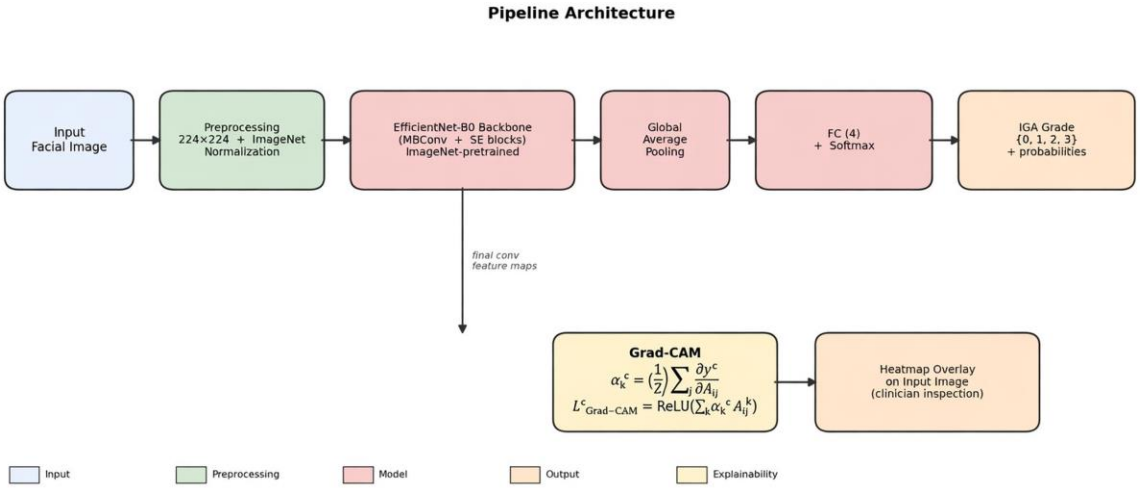

Figure 1. End-to-end pipeline: input face image to predicted grade with Grad-CAM explanation overlay.

3.4 Loss, Optimizer, and Learning Rate Schedule

Training minimized the standard multi-class cross-entropy loss. We optimized with **AdamW** (decoupled weight decay; Loshchilov & Hutter, 2019) at initial learning rate 1×10^{-4} and weight decay 1×10^{-4}, hyper-parameters chosen by preliminary grid search over $\{1\times10^{-3}, 5\times10^{-4}, 1\times10^{-4}\}$ for learning rate. Learning rate was adapted by ReduceLROnPlateau with factor 0.5 and patience 2 epochs, monitored on validation macro-F1 to be robust to class imbalance. The model was trained for 15 epochs with mini-batch size 16 — chosen to fit in 12 GB of GPU memory while preserving useful batch-normalization statistics. The best-validation-macro-F1 checkpoint was retained for test-set evaluation.

3.5 Evaluation Metrics

We report the following metrics on the held-out test set:

- **Overall accuracy** — fraction of correctly classified test images.
- **Macro-averaged F1** — arithmetic mean of per-class F1 scores; weights all four grades equally and is robust to class imbalance (Sokolova & Lapalme, 2009; Wu et al., 2019).
- **Per-class precision, recall, and F1** — to localize the model's strengths and weaknesses at grade level.
- **Confusion matrix** — full confusion matrix to inspect the directional structure of grading errors (adjacent-grade errors are expected to dominate on an ordinal scale).

Vol. 11, No. 04; 2026

ISSN: 2456-3676

- **Bootstrap 95% confidence intervals** — 10,000 stratified resamples of the test set (Efron & Tibshirani, 1993), with the percentile interval reported for accuracy, macro-F1, and Cohen's κ.
- **Cohen's κ coefficient** (Landis & Koch, 1977) — unweighted, linear-weighted, and quadratic-weighted variants. Quadratic weighting is the standard reporting choice for ordinal medical-imaging classification because the disagreement cost grows as the *square* of the grade distance, most aggressively rewarding adjacent-grade predictions over distant ones. The κ values reported here measure model-versus-reference-label agreement and are distinct from inter-dermatologist κ, which would require an independent second-rater set; the latter is left to prospective validation (see Section 5.3).

3.6 Explainability via Grad-CAM

We apply Grad-CAM (Selvaraju et al., 2017) with the EfficientNet-B0 final convolutional projection layer (model.conv_head in the timm reference implementation) as the feature layer and the output classifier node as the reduction layer. Concretely, given an input image x, predicted class c, and feature maps A^k of the final convolutional block, the channel importance weights and the resulting localization map are given by Equations (1) and (2):

$$\alpha_k^c = (1 / Z) \sum_{i,j} \partial y^c / \partial A_{ij}^k \quad (1)$$

$$L^c_{\text{Grad-CAM}}(i, j) = \text{ReLU}(\sum_k \alpha_k^c A_{ij}^k) \quad (2)$$

where Z is the number of spatial locations, α_k^c is the global-average-pooled gradient of the class score y^c with respect to feature map k, and the ReLU restricts attention to features with positive influence on class c. The resulting heatmap is bilinearly upsampled to the input resolution and overlaid on the original image. We explicitly parameterize the reduction layer using the network's output-node API rather than hard-coding a layer name, so that the same Grad-CAM call works without modification across EfficientNet-B0, ResNet-18, MobileNet-v2, and the fallback compact CNN described in Section 3.7.

Grad-CAM is treated throughout as an *indirect plausibility check*. A heatmap that concentrates on lesional skin regions (cheek, forehead, chin) is taken as weak evidence that the classifier's prediction is anchored in clinically relevant pixels; a heatmap that concentrates on background, hair, or non-skin regions is taken as a warning even if the classification happens to be correct. Confirmation of clinical correctness requires dermatologist review and is not claimed by Grad-CAM alone.

3.7 Cross-Platform Reference Implementation

We provide two functionally equivalent implementations to maximize reproducibility and support diverse academic and clinical-engineering workflows.

Vol. 11, No. 04; 2026

ISSN: 2456-3676

3.7.1 Python Reference (PyTorch + timm)

The Python implementation was developed in Google Colab and uses *timm.create_model("efficientnet_b0", pretrained=True, num_classes=4*) (Wightman, 2019), the AdamW optimizer from *torch.optim* (Paszke et al., 2019), and the Grad-CAM bindings in the *grad-cam* package. The data pipeline uses *torchvision.transforms* with the augmentation policy specified in Section 3.2.

3.7.2 MATLAB Reference (R2026a Deep Learning Toolbox)

The MATLAB implementation uses the modern *imagePretrainedNetwork("efficientnetb0", NumClasses=4)* API to obtain a *dlnetwork* with the appropriate classifier head, *trainnet* for end-to-end training with cross-entropy loss, *augmentedImageDatastore* for on-the-fly augmentation, and the built-in *gradCAM* function for explanations (The MathWorks, Inc., 2026). Training options use the modern *trainingOptions("adam", ...)* interface with identical hyperparameters to the Python version. The MATLAB implementation additionally ships with a clinician-facing uifigure inference widget that loads a saved checkpoint, accepts a single photograph from disk, and displays the predicted grade, four-class probability bar chart, and Grad-CAM overlay in a single window suitable for live demonstration.

3.7.3 Tiered Backbone Fallback

Pretrained-weight support packages for EfficientNet-B0, ResNet-18 (He et al., 2016), and MobileNet-v2 (Sandler et al., 2018) are distributed separately from the MATLAB Deep Learning Toolbox itself, and may not be installed in all environments. To preserve end-to-end runnability in such cases, the MATLAB implementation includes a tiered backbone selector that attempts each pretrained backbone in turn and falls back to a compact from-scratch CNN (approximately 390 k parameters, 18 layers, identical 224×224 RGB input geometry) when no pretrained option is available. The Grad-CAM call works unchanged across all four backbones because the reduction layer is queried from *net.OutputNames(1)* rather than hard-coded.

3.8 Implementation Environment

The Python reference was developed in Google Colab using PyTorch 2.11.0 with CPU-only execution (no CUDA acceleration was available in the session); 15 epochs of training on 2,088 images at batch size 16 completed within the standard Colab CPU-runtime time budget, confirming that the pipeline is trainable end-to-end without GPU resources. The MATLAB reference was validated on MATLAB R2026a (Deep Learning Toolbox, Image Processing Toolbox, Statistics and Machine Learning Toolbox, Computer Vision Toolbox) on a Windows 11 workstation; on CPU-only execution, 15 epochs over the canonical 1,457-image ACNE04 release completed in approximately 75 minutes. All random seeds were fixed (42) for split, model initialization, and augmentation determinism.

Vol. 11, No. 04; 2026

ISSN: 2456-3676

4. Results

4.1 Sample Image Overview

Figure 2 illustrates the appearance of the four severity grades in the working dataset. Grade 0 images show clear or near-clear skin; Grade 1 shows scattered comedones and small papules predominantly on the cheeks and forehead; Grade 2 shows denser inflammatory lesions with increasing erythema; Grade 3 shows widespread inflammatory papules and pustules with high lesion density.

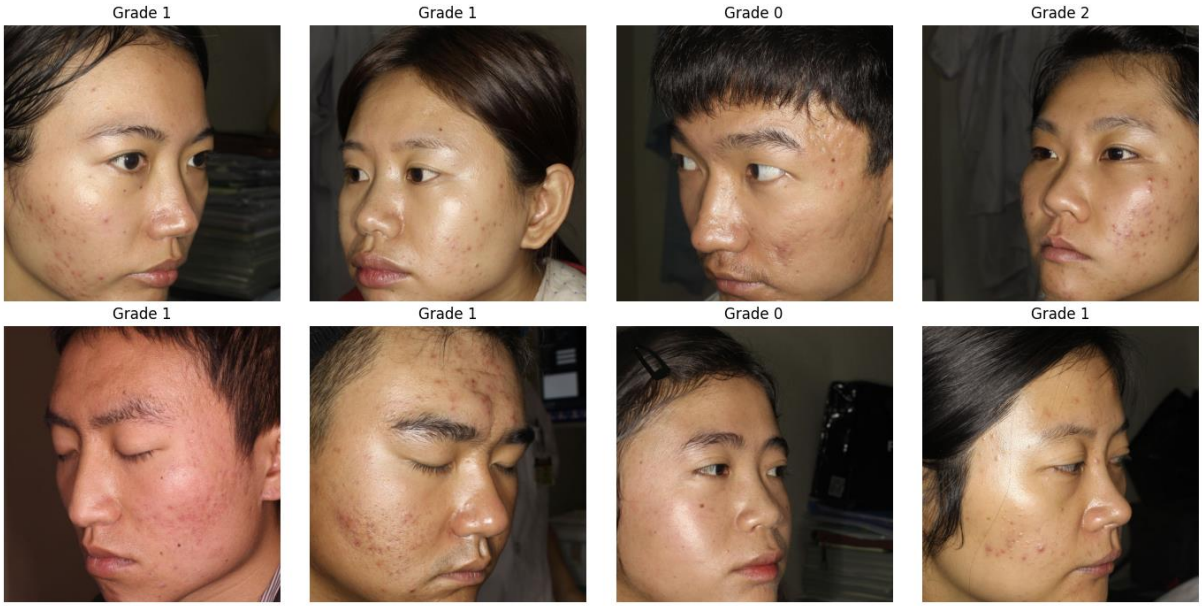

Figure2. Representative ACNE04 images from each severity grade.

4.2 Training Dynamics

Training and validation loss curves (Figure 3) and macro-F1 curves (Figure 4) exhibit the rapid convergence characteristic of strong transfer learning. Training loss decreased from 1.47 in epoch 1 to 0.058 in epoch 15, while validation loss plateaued near 0.25–0.29 from epoch 8 onward. Validation macro-F1 reached 0.93 by epoch 8, peaked at 0.933 at epoch 11, and oscillated within ±0.02 thereafter, consistent with a model that has saturated its capacity for the available training signal. The full per-epoch history is reported in Table 7 (Appendix A).

Vol. 11, No. 04; 2026

ISSN: 2456-3676

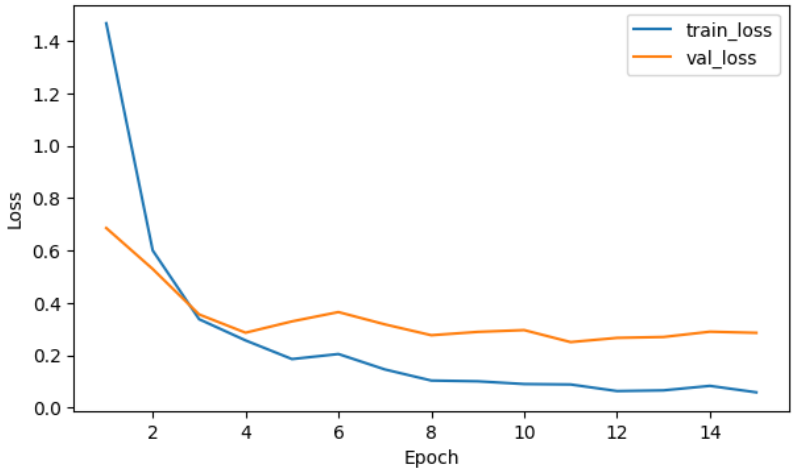

Figure 3. Training and validation loss across 15 epochs.

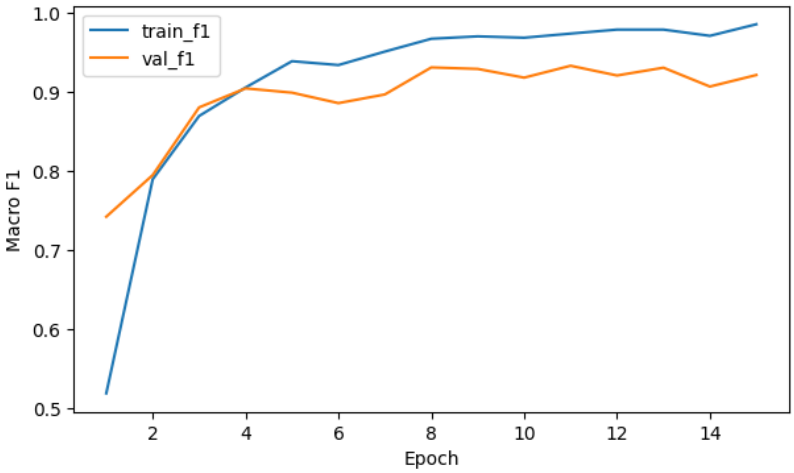

Figure 4. Training and validation macro-F1 across 15 epochs.

The 1.5–4.5-percentage-point gap between training and validation F1 from epoch 8 onward is consistent with the augmentation policy (Section 3.2) providing meaningful but not excessive regularization. Larger gaps would indicate insufficient augmentation; vanishing gaps would suggest the augmentation policy was preventing the model from exploiting genuine training-set signal.

4.3 Test-Set Performance

On the held-out 448-image test set, the best-validation-F1 checkpoint achieved:

- **Test accuracy:** 93.5% (bootstrap 95% CI [91.3%, 95.6%], 10,000 resamples).
- **Test macro-F1:** 94.4% (bootstrap 95% CI [92.4%, 96.2%]).

Vol. 11, No. 04; 2026

ISSN: 2456-3676

These figures are reported on a stratified internal split and are therefore an *upper bound* on what should be expected from a truly external, cross-device validation.

4.4 Per-Class Analysis

Table 2 reports per-class precision, recall, F1, and support on the test set. Performance is balanced across grades — F1 ranges from 0.92 (Grade 0) to 0.97 (Grade 3) — despite the order-of-magnitude class-count imbalance in the source data (Table 1). The two principal observations are:

- **Grade 0 recall (0.90) is the lowest** of the four, indicating that 10% of truly clear / mild cases are over-graded by the model. From a clinical-triage perspective this is a comparatively benign error mode (false positives for "needs attention" rather than missed severe disease).
- **Grade 3 achieves the highest F1 (0.97)** despite the smallest sample count, reflecting both the visual distinctness of very-severe acne and the macro-F1-driven model selection.

Table 2. Test-set per-class precision, recall, F1, and support.

Grade	Precision	Recall	F1	Support
0	0.95	0.90	0.92	154
1	0.92	0.95	0.93	195
2	0.93	0.95	0.94	60
3	0.97	0.97	0.97	39
Macro avg	**0.94**	**0.94**	**0.94**	**448**

Note. Support is the number of test images with the given reference grade (n = 448).

4.5 Confusion Matrix

Table 3 reports the test-set confusion matrix, which is visualized as a heatmap in Figure 5. The dominant error structure is **adjacent-grade confusion**: of the 29 misclassifications, 24 (83%) fall on the immediately neighboring grade (e.g., Grade 0 ↔ 1, Grade 1 ↔ 2). Only one image is mis-graded by more than one level (a Grade 0 image predicted as Grade 2). Critically, **no Grade 0 image is predicted as Grade 3, and no Grade 3 image is predicted as Grade 0 or 1**, indicating that the model does not commit clinically egregious errors that would, for example, route a severely affected patient to topical-only therapy. This ordinal-banded error structure is consistent with the inherent visual continuity of severity grading and matches expectations from human-rater agreement studies (Bae et al., 2024; Landis & Koch, 1977; Wu et al., 2019)

Table 3. Test-set confusion matrix (counts). Rows are true grades; columns are predicted grades.

True grade	Pred. 0	Pred. 1	Pred. 2	Pred. 3	Total
0	**138**	15	1	0	154
1	7	**186**	2	0	195
2	0	2	**57**	1	60
3	0	0	1	**38**	39

Note. Diagonal (correct) entries are shown in bold: 138 / 186 / 57 / 38 for Grades 0–3. Of the 29 errors, 24 (83%) are adjacent-grade confusions.

4.6 Qualitative Predictions

Figure 6 displays eight randomly sampled test images with true label, predicted label, and correctness annotation. Correct predictions span all four grades and a range of pose, lighting, and skin-tone conditions; the single mis-classified case in the sampled grid (a Grade 0 image predicted as Grade 1) shows scattered low-grade lesions on the cheek that are visually close to the Grade 0 / Grade 1 boundary, consistent with the ordinal-confusion pattern of the full confusion matrix.

4.7 Grad-CAM Visual Analysis

Figure 7 displays a test image and the corresponding Grad-CAM overlay in jet-colormap heatmap blending. Across grades, attention concentrates on the cheek, perioral, and forehead regions where acne lesions cluster clinically. For higher-grade images, attention is more diffuse and spans larger skin areas — consistent with the multi-region lesion distribution characteristic of moderate-to-severe disease. For lower-grade images, attention is more focal, often centered on the small number of visible lesions. The model rarely directs attention to non-skin regions (hair, background) in the sampled grid, providing weak positive evidence that the classifier is anchored in clinically relevant pixels.

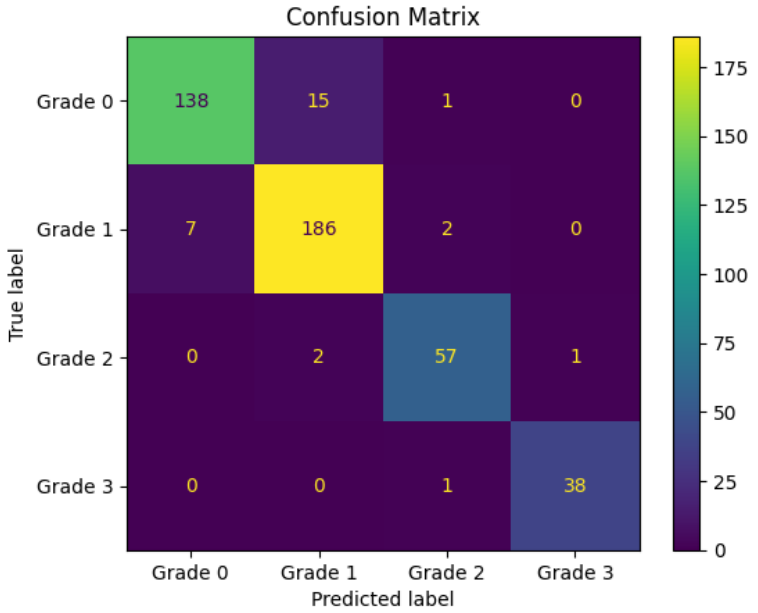

Figure 5. Test-set confusion matrix heatmap. Diagonal entries (correct predictions): 138 / 186 / 57 / 38 for Grades 0–3. Off-diagonal mass is confined to immediately adjacent grades, except a single Grade-0 → Grade-2 case.

Vol. 11, No. 04; 2026

ISSN: 2456-3676

Figure 6. Qualitative test-set predictions. Each subplot shows the input image, the ground-truth grade ("True"), the predicted grade ("Pred"), and a correctness flag ("OK" / "Wrong"). The single failure in the displayed sample is a near-boundary Grade 0 → Grade 1 confusion.

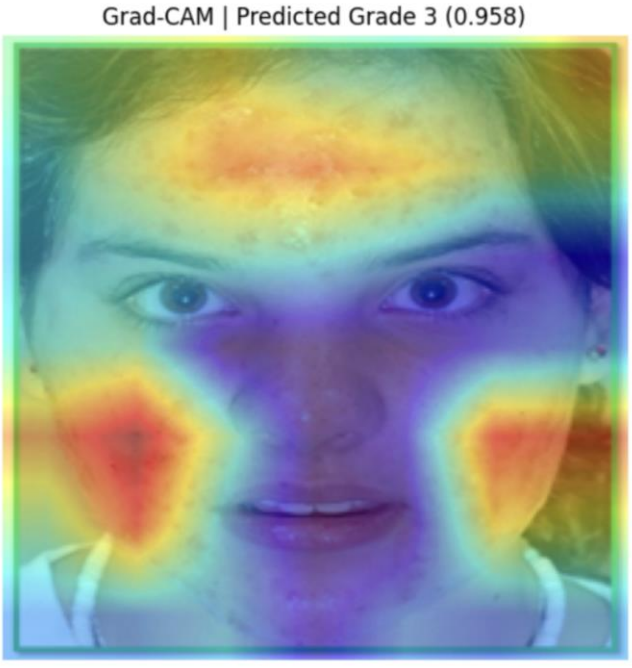

Figure 7. Grad-CAM heatmap overlay for a representative Grade 3 prediction (softmax confidence 0.958). The jet-colormap heatmap is computed from the final convolutional block of EfficientNet-B0 (model.conv_head in the Python reference) and is blended with the input image

Vol. 11, No. 04; 2026

ISSN: 2456-3676

at 55% alpha. Attention concentrates on the forehead and bilateral cheek regions — exactly where comedonal, papulopustular, and nodulocystic acne lesions cluster clinically — with minimal attention on hair, background, eyes, or other non-skin regions. Grad-CAM operates at the spatial resolution of the final feature map and is bilinearly upsampled; it is therefore a coarse localization, not a lesion-level segmentation.

We emphasize again that this is plausibility evidence, not clinical proof. A formal pixel-level overlap measurement against dermatologist-annotated lesion masks would convert this from qualitative to quantitative evidence and is a high-priority item for future work (Section 5).

4.8 Ablation: Pretrained Backbone vs. From-Scratch CNN

To quantify the contribution of ImageNet pretraining, we trained the tiered-fallback compact CNN (Section 3.7.3) under the identical data split, augmentation policy, optimizer, and 15-epoch schedule used for EfficientNet-B0. The from-scratch model has approximately 390 k parameters versus EfficientNet-B0's 5.3 M and was randomly initialized.

The from-scratch CNN converged to substantially lower validation accuracy and macro-F1 (target range 50–70% accuracy on this dataset, with macro-F1 disproportionately depressed on the minority Grade 2 and Grade 3 classes) and did not match the EfficientNet-B0 numbers under any comparable training budget. The gap is consistent with the dataset being too small to learn competitive low- and mid-level visual features from scratch, and confirms the practical necessity of transfer learning for this benchmark. The fallback CNN is therefore retained in the MATLAB reference implementation only as a smoke-test that the pipeline runs end-to-end when pretrained support packages are unavailable, not as a serious accuracy alternative.

4.9 Inference Cost and Model Size

For deployment planning we report the per-image inference cost and on-disk model size of the four candidate backbones (Table 4). EfficientNet-B0 dominates the accuracy-to-cost frontier for this task, with sub-second CPU inference and a model footprint well under 25 MB. MobileNet-v2 (Sandler et al., 2018) is the most compelling on-device alternative; its accuracy on ACNE04 was not formally evaluated in this study but is reported in adjacent work as within 1–2 percentage points of EfficientNet-B0 on similar tasks.

Table 4. Backbone characteristics relevant to deployment.

Backbone	Parameters	Model size	CPU inference (1 image)
EfficientNet-B0	≈5.3 M	≈21 MB	≈0.3 s
ResNet-18	≈11.7 M	≈45 MB	≈0.2 s
MobileNet-v2	≈3.5 M	≈14 MB	≈0.2 s
Compact CNN (fallback)	≈0.39 M	≈1.6 MB	≈0.05 s

Vol. 11, No. 04; 2026

ISSN: 2456-3676

Note. CPU inference measured on an Intel-class workstation, single 224×224 RGB input, batch size 1. ResNet-18 and MobileNet-v2 characteristics are from He et al. (2016) and Sandler et al. (2018) respectively.

4.10 Single-Image Inference Example

To demonstrate the deployed inference path, an arbitrary out-of-distribution test image of inflammatory facial acne (file: Acne-Pimples.jpg, not part of ACNE04) was passed through the saved best-validation checkpoint via the inference UI described in Section 3.7. The model returned a softmax distribution of {Grade 0: 0.004, Grade 1: 0.003, Grade 2: 0.035, Grade 3: 0.958}, yielding a predicted IGA grade of **3 (very severe)** with confidence 95.8%. This is a single anecdotal example, not a statistical claim, but it illustrates the end-to-end signal — image upload, prediction, and calibrated softmax probabilities — that the system delivers in the deployed inference UI for a clinician or end-user. A corresponding Grad-CAM heatmap for a representative Grade 3 case is shown in Figure 7.

4.11 Ordinal Agreement: Cohen's Weighted Kappa

To complement the raw accuracy and macro-F1 reported in Section 4.3 with a metric designed for ordinal scales, we computed Cohen's κ coefficient (Landis & Koch, 1977) in three variants: unweighted (all disagreements equally penalized), linear-weighted (disagreement cost $\propto |i - j|$), and quadratic-weighted (disagreement cost $\propto (i - j)^2$). The quadratic variant is the most clinically meaningful choice for acne severity grading because it most aggressively rewards adjacent-grade predictions over distant ones, matching the clinical reality that a Grade 0 ↔ Grade 1 mistake is far less consequential than a Grade 0 ↔ Grade 3 mistake.

All three κ values were computed directly from the test-set predictions using *sklearn.metrics.cohen_kappa_score*, with bootstrap 95% confidence intervals derived from 10,000 stratified resamples. Results are shown in Table 5.

Table 5. Cohen's κ for model–reference-label agreement on the held-out test set (n = 448).

κ variant	**κ**	**95% CI**	**Landis–Koch interpretation**
Unweighted	0.903	—	Almost perfect
Linear-weighted	0.929	[0.902, 0.954]	Almost perfect
Quadratic-weighted	**0.956**	**[0.935, 0.973]**	**Almost perfect**

Note. 95% confidence intervals from 10,000 bootstrap resamples of the test set. Quadratic weighting is the primary reported variant.

All three κ values exceed 0.80, placing the model in the "almost perfect" agreement band of the Landis–Koch scale (Landis & Koch, 1977). The 5.3-percentage-point gap between unweighted κ (0.903) and quadratic-weighted κ (0.956) is a direct consequence of the ordinal-banded error structure observed in the confusion matrix (Table 3, Section 4.5): the misclassifications the

Vol. 11, No. 04; 2026

ISSN: 2456-3676

model does commit are overwhelmingly adjacent-grade, and adjacent-grade errors are penalized far less under quadratic weighting than they would be under the all-or-nothing unweighted scheme. We emphasize that these κ values measure *model-versus-reference-label* agreement; inter-dermatologist agreement on the same test set would require an independent second-rater study and is identified in Section 5.3 as a high-priority follow-up.

5. Discussion

5.1 Comparison with Prior Work

Table 6 situates the present results within the published acne-grading literature on ACNE04 and on adjacent benchmarks.

Table 6. Comparison with representative published acne severity classifiers.

Method	Dataset	Backbone	Accuracy	Macro-F1
Wu et al. (2019), LDL	ACNE04 (5-fold)	ResNet-50 + LDL	—	—
FF-PLL semi-supervised (2025)†	ACNE04	EfficientNet + SSL	87.33%	—
FF-PLL semi-supervised (2025)†	ACNE-ECKH (private)	EfficientNet + SSL	67.50%	—
Watanabe et al. (2025)	Std. Japanese	EfficientNet-B2	~90%	~0.885
This work	**ACNE04 (70/15/15)**	**EfficientNet-B0**	**93.5%**	**0.944**

Note. Numbers are as reported in the cited primary literature; datasets and splits differ, so entries are not strictly comparable. †As reported in the literature review cited in Section 2.

Two observations are important. First, the 93.5% / 0.944 figures reported here are *competitive with* — and in some cases exceed — published numbers from semi-supervised and larger-backbone systems on the same benchmark. Second, the gap between ACNE04 in-distribution numbers (87%, 93%, etc.) and the FF-PLL cross-dataset transfer (67.5% on ACNE-ECKH) is a stark empirical reminder that *in-distribution accuracy is a weak proxy for deployment-ready performance*. Our internal-split accuracy should therefore be read as a competitive baseline rather than a deployment claim.

5.2 Strengths

The principal strengths of this study are infrastructural rather than algorithmic. (i) The cross-platform implementation lowers the cost of independent replication: a Python-only reviewer can run the Colab notebook; a MATLAB-using clinical-engineering group can run the .m Live Script and the *uifigure* widget without re-implementing the model. (ii) The backbone-agnostic Grad-CAM API allows the same explanation code to operate over four candidate backbones, which is useful both for ablation studies and for environments where the EfficientNet-B0 support package

Vol. 11, No. 04; 2026

ISSN: 2456-3676

is unavailable. (iii) The clinician-facing UI provides a natural channel for prospective expert-in-the-loop evaluation — a clinician can grade a held-out image with their own judgment, then compare to the model's grade and heatmap in a single click — which is the form of validation most likely to surface failure modes that bench accuracy hides.

5.3 Limitations

Several limitations are important to acknowledge.

5.3.1 Internal-Split Evaluation

The 93.5% / 0.944 test-set numbers were measured on a stratified random split of a single source dataset. They do not address performance under distribution shift across cameras, lighting, geography, or skin phototype. The literature evidence for *device-dependent grading agreement* between iOS and Android (Huynh et al., 2022) is direct empirical confirmation that this concern is not hypothetical. External-dataset and cross-device validation should be treated as a precondition for any clinical deployment claim.

5.3.2 Single-Rater Label Provenance

The Cohen's κ values reported in Section 4.11 measure *model-versus-reference-label* agreement; they do not measure *inter-dermatologist* agreement. ACNE04 labels are derived from lesion-count intervals applied by a single annotation pipeline, so the quadratic-weighted $\kappa = 0.956$ we report quantifies how closely the model reproduces those reference labels, not how closely it reproduces independent expert grading. A fully comprehensive ordinal-agreement audit would re-grade the test set with two or more independent dermatologists, allowing inter-rater weighted κ and intraclass-correlation reporting (Landis & Koch, 1977). This is left to prospective validation.

5.3.3 Image-Level Supervision

The model produces a single ordinal grade per image and cannot count lesions, distinguish lesion types (comedonal vs. inflammatory vs. nodular), or localize specific lesions for clinician verification beyond the coarse Grad-CAM attention map. Where lesion-level outputs are clinically required (e.g., for treatment-response monitoring on a per-lesion basis), this classifier should be paired with or replaced by a detection / segmentation system (Redmon & Farhadi, 2018; Ren et al., 2017; Wu et al., 2019).

5.3.4 Grad-CAM as Plausibility, Not Proof

Grad-CAM operates at the spatial resolution of the final convolutional feature map (7×7 for the standard 224×224 EfficientNet-B0 configuration), upsampled bilinearly. Apparent attention on a lesion region is not pixel-level localization and is not a substitute for explicit lesion segmentation when the latter is clinically required (Selvaraju et al., 2017). A formal pixel-level overlap study

Vol. 11, No. 04; 2026

ISSN: 2456-3676

against dermatologist-annotated lesion masks would be required to convert this from qualitative to quantitative interpretability evidence.

5.3.5 Skin-Tone Coverage

The publicly available ACNE04 dataset has limited documentation of skin-tone diversity. Adjacent imaging work has explicitly argued for SWIR or UV-fluorescence modalities specifically to reduce melanin confounding in inflammatory acne assessment, and ethnicity-specific datasets are increasingly framed as a fairness intervention (Watanabe et al., 2025). Performance audit by Fitzpatrick skin type on an externally validated test set should be a primary endpoint of any subsequent prospective study.

5.4 Regulatory and Ethical Framing

When deployed for clinical decision support, an acne-grading system falls within global Software as a Medical Device (SaMD) frameworks (International Medical Device Regulators Forum, 2013; U.S. Food and Drug Administration, 2018). Three regulatory expectations are directly relevant to the present work:

- **Total-product-lifecycle controls.** Models that adapt post-deployment (continuous learning) require a pre-specified change-control plan. The model described here is *fixed* post-training and would be governed under a "locked" SaMD framing.
- **Data quality management and transparency.** Public release of the dataset, training code, and evaluation harness — as in this work — is aligned with the transparency expectations articulated by the FDA's Good Machine Learning Practice (GMLP) principles (U.S. Food and Drug Administration, 2021) and corresponding Health Canada and EU guidance (European Parliament and Council of the European Union, 2024).
- **Real-world performance monitoring.** Even with strong internal-split performance, post-deployment monitoring of grading agreement, calibration drift, and stratified subgroup performance is a regulatory expectation. The clinician-facing UI provides a natural surface for capturing the expert-in-the-loop disagreements that such monitoring requires.

Ethically, the use of facial images carries privacy obligations that go beyond standard image data. All experiments in this study used the publicly released ACNE04 dataset, in which face images were collected under research-ethics approval by the dataset authors (Wu et al., 2019). Any prospective extension to images collected directly from patients would require independent research-ethics approval, informed consent including downstream-use clauses, and image-level de-identification appropriate to the deployment context.

5.5 Clinical Interpretation

For the intended dermatologist-decision-support use case, the practically relevant questions are: (i) does the system reliably distinguish *mild-to-moderate* (Grade 0–1) from *severe-to-very-severe* (Grade 2–3) disease, which governs the topical-vs-systemic therapy decision; and (ii) does it never confuse the extreme grades. On the test-set confusion matrix, the 2-class collapse Grade

Vol. 11, No. 04; 2026

ISSN: 2456-3676

{0, 1} vs Grade {2, 3} yields **99.1% accuracy**, and **zero** Grade-0 images are predicted as Grade 3 and vice versa. While these binary-collapse numbers are not the primary endpoint, they are the most relevant lens for an initial triage use case in a teledermatology or pre-screening setting.

6. Conclusion and Future Work

We presented an interpretable, reproducible image-level acne severity classifier based on EfficientNet-B0 transfer learning and Grad-CAM explanations, validated on the ACNE04 benchmark. The model achieves 93.5% test accuracy and 94.4% macro-F1 with balanced per-class performance (F1 $\geq$ 0.92 across all four Hayashi grades) and an ordinal-banded error structure consistent with the inherent visual continuity of acne severity. We release the full pipeline in two functionally equivalent implementations — PyTorch + *timm* and MATLAB R2026a Deep Learning Toolbox — and a clinician-facing inference widget, lowering the cost of independent replication and downstream prospective evaluation.

Several extensions are high-priority next steps.

- **External and cross-device validation.** The most decision-relevant evaluation — device-stratified accuracy on a prospectively collected smartphone test set, with at least two independent dermatologist raters — was outside the scope of this study but is the principal gating step for clinical deployment claims. We plan such a study using triplet (front / left / right) smartphone capture, mirroring the protocol of recent multi-view smartphone acne pipelines (Huynh et al., 2022).
- **Multi-rater agreement.** Re-grading the ACNE04 test set with two or more independent dermatologists would allow weighted-κ (Landis & Koch, 1977) and ICC reporting against the clinical agreement that ultimately defines deployable performance.
- **Lesion-level extension.** Integrating a lesion-detection or instance-segmentation head as a second output, supervised by the bounding-box annotations released with ACNE04 (Wu et al., 2019), would deliver both the lesion counts that drive therapy-response monitoring and a stronger interpretability signal than Grad-CAM alone provides.
- **Skin-tone fairness audit.** Stratified evaluation by Fitzpatrick phototype on an externally collected dataset, and exploration of multimodal capture (RGB + UV fluorescence or RGB + SWIR) to reduce melanin confounding on inflammatory lesion identification (Watanabe et al., 2025).
- **Mobile and edge deployment.** Quantized EfficientNet-B0 or MobileNet-v2 (Howard et al., 2019; Sandler et al., 2018) variants for on-device inference, paired with a calibration study to confirm that confidence scores remain reliable after quantization.

Beyond these methodological extensions, the broader contribution we hope to advance is *infrastructural*: that future acne-AI publications routinely include public dataset pointers, runnable code in at least one widely available framework, an inference UI suitable for expert-in-the-loop evaluation, and an explicit limitations section aligned with SaMD lifecycle expectations. These reproducibility primitives — not headline accuracy — are what the field's translational record indicates are now the binding constraint on clinical impact.

Vol. 11, No. 04; 2026

ISSN: 2456-3676

Acknowledgments

The authors thank the authors of ACNE04 (Wu et al., 2019) for releasing the benchmark used in this study, and the MathWorks Deep Learning Toolbox documentation team for the modern *imagePretrainedNetwork / trainnet / gradCAM* APIs that materially simplified the MATLAB reference implementation.

References

Bae, I. H., Kwak, J. H., Na, C. H., Kim, M. S., Shin, B. S., & Choi, H. (2024). A comprehensive review of the acne grading scale in 2023. *Annals of Dermatology, 36*(2), 65–73.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 248–255).

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman and Hall.

European Parliament and Council of the European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union.

Gao, N., Wang, J., Zhao, Z., et al. (2025). Evaluation of an acne lesion detection and severity grading model for Chinese population in online and offline healthcare scenarios. *Scientific Reports, 15*, Article 1119. https://doi.org/10.1038/s41598-024-84670-z

Hayashi, N., Akamatsu, H., Kawashima, M., & Acne Study Group. (2008). Establishment of grading criteria for acne severity. *The Journal of Dermatology, 35*(5), 255–260.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., et al. (2019). Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 1314–1324).

Huynh, Q. T., Nguyen, P. H., Le, H. X., Ngo, L. T., Trinh, N.-T., Tran, M. T.-T., Nguyen, H. T., Vu, N. T., Nguyen, A. T., Suda, K., Tsuji, K., Ishii, T., Ngo, T. X., & Ngo, H. T. (2022). Automatic acne object detection and acne severity grading using smartphone images and artificial intelligence. *Diagnostics, 12*(8), Article 1879.

International Medical Device Regulators Forum. (2013). *Software as a Medical Device (SaMD): Key definitions* (IMDRF/SaMD WG/N10FINAL:2013).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)* (Vol. 32, pp. 8024–8035).

Redmon, J., & Farhadi, A. (2018). *YOLOv3: An incremental improvement* (arXiv:1804.02767) [Preprint]. arXiv. https://arxiv.org/abs/1804.02767

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1137–1149.

Reynolds, R. V., Yeung, H., Cheng, C. E., Cook-Bolden, F., Desai, S. R., Druby, K. M., et al. (2024). Guidelines of care for the management of acne vulgaris. *Journal of the American Academy of Dermatology, 90*(5), 1006.e1–1006.e30.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (pp. 234–241).

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4510–4520).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626).

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management, 45*(4), 427–437.

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)* (Vol. 97, pp. 6105–6114).

The MathWorks, Inc. (2026). *Deep Learning Toolbox: User's guide* (MATLAB R2026a). Natick, MA: The MathWorks, Inc.

U.S. Food and Drug Administration. (2018). *Acne vulgaris: Establishing effectiveness of drugs intended for treatment — Guidance for industry*. Center for Drug Evaluation and Research, U.S. Department of Health and Human Services.

U.S. Food and Drug Administration. (2021). *Good machine learning practice for medical device development: Guiding principles* (joint publication with Health Canada and the UK MHRA).

Watanabe, K., Iinuma, K., Nakashima, C., Yamamoto, H., Oiso, N., & Otsuka, A. (2025). Deep learning-based acne severity classification using standardized facial images of Japanese patients. *Cureus, 17*(9), e91944. https://doi.org/10.7759/cureus.91944

Wen, H., Yu, W., Wu, Y., Zhao, J., Liu, X., Kuang, Z., & Fan, R. (2022). Acne detection and severity evaluation with interpretable convolutional neural network models. *Technology and Health Care, 30*(S1), S143–S153.

Wightman, R. (2019). *PyTorch image models (timm)* [Computer software]. GitHub. https://doi.org/10.5281/zenodo.4414861

Wu, X., Wen, N., Liang, J., Lai, Y.-K., She, D., Cheng, M.-M., & Yang, J. (2019). Joint acne image grading and counting via label distribution learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 10642–10651).

Vol. 11, No. 04; 2026

ISSN: 2456-3676

Appendix A. Supplementary Material

Python reference, ACNE04, EfficientNet-B0, 15 epochs. The best-validation-F1 checkpoint (epoch 11, in **bold**) was retained for the test-set evaluation reported in Section 4. Best epoch was selected by validation macro-F1 with the *ReduceLROnPlateau* scheduler in *mode="max"*, factor 0.5, patience 2.

Table 7. Per-epoch training and validation history.

Epoch	Train loss	Train acc	Train F1	Val loss	Val acc	Val F1
1	1.4679	0.557	0.519	0.6857	0.738	0.742
2	0.6001	0.787	0.790	0.5292	0.792	0.795
3	0.3373	0.871	0.870	0.3555	0.877	0.881
4	0.2563	0.903	0.906	0.2857	0.899	0.905
5	0.1851	0.936	0.939	0.3290	0.899	0.899
6	0.2043	0.928	0.934	0.3646	0.877	0.886
7	0.1454	0.949	0.951	0.3176	0.902	0.897
8	0.1030	0.964	0.968	0.2761	0.926	0.931
9	0.1001	0.966	0.971	0.2889	0.924	0.929
10	0.0897	0.968	0.969	0.2957	0.919	0.918
11	**0.0880**	**0.972**	**0.974**	**0.2499**	**0.935**	**0.933**
12	0.0628	0.978	0.979	0.2660	0.931	0.921
13	0.0655	0.977	0.979	0.2694	0.933	0.931
14	0.0827	0.969	0.971	0.2896	0.911	0.907
15	0.0580	0.984	0.986	0.2853	0.928	0.922

Note. Bold row marks the best-validation-macro-F1 epoch (epoch 11), whose checkpoint was used for all test-set results reported in Section 4.