DOI:doi.org/10.65766/alyj.2026.24.01.13

Lossless Float16 Quantization for Multi-Class Skin Lesion Classification: Lesion-Level Stratified Evaluation on ISIC 2019

Haitham Qutaiba Ghadhban
Department of Computer Engineering – Faculty of Engineering – University of Diyala –Iraq
Corresponding Author Email:haithamqutaiba@uodiyala.edu.iq

Important Dates
Received: 5/4/2026, Accepted: 9/5/2026, Published: 30/6/2026

Abstract

Image-level random data partitioning, which continues to plague the field through intra-lesion data leakage training and test sets may both contain multiple images of the same lesion is still the dominant methodological limitation in dermoscopic AI research that produces falsely high but unacceptable metrics. However, the impacts of this leakage on Post-Training Quantization (PTQ) benchmarking have not yet been investigated. This work proposes a leakage-free, precise lesion-level stratified evaluation framework for multi-class skin lesion classifier evaluating PTQ. A total of 23,247 images from eight diagnostic classes of ISIC 2019 benchmark were divided based on 11,847 unique lesions. Our two-stage EfficientNetB0 transfer learning pipeline with the focal loss which can down weight easy-to-classify samples achieves 86.98% macro-AUC and 84.05% top-2 accuracy. Then, utilizing Float16 PTQ for further model size reduction (49.6% compression ratio all models performance degradation is not statistically significant), a lossless compression through vigorous lesion-level evaluation is demonstrated. Variance analysis on three independent splits shows that data partitioning alone can explain up to ±4.28% top-1 accuracy variability, as opposed to quantization precision which matter barely less.

Keywords

Class imbalance, Dermoscopic image analysis, Model compression, Post-training quantization, Transfer learning.

لتصنيف آفات الجلد متعددة الفئات: تقييم طبقي على Float16 التكميم الخالي من الفقدان بدقة مستوى الآفة باستخدام قاعدة بيانات ISIC 2019

هيثم قتيبة غضبان

قسم هندسة الحاسوب، كلية الهندسة، جامعة ديالى

ايميل الباحث المراسل: haithamqutaiba@uodiyala.edu.iq

DOI:doi.org/10.65766/alyj.2026.24.01.13

الملخص

تتجه أنظمة تصنيف الصور التشخيصية بالمنظار الجلدي بشكل متزايد نحو النشر على الأجهزة الطرفية محدودة الموارد، حيث تبقى ضغط النماذج وشفافيتها وسلامة التقييم من المتطلبات الجوهرية. يتمثل القصور المنهجي السائد في الأدبيات العلمية في اعتماد التقسيم العشوائي على مستوى الصور، مما يتيح ظهور صور متعددة للآفة ذاتها في مجموعتي التدريب والاختبار في آن واحد. حين لا يعالج هذا الخلل، يفضي التسرب داخل الآفة الواحدة إلى تضخيم مقاييس التقييم المبلغ عنها، إذ تستغل النماذج الخصائص البصرية الخاصة بالمريض بدلا من تعلم الأنماط السريرية القابلة للتعميم. علاوة على ذلك، لم تستكشف حتى الآن تداعيات هذا التسرب على دراسات قياس أداء التكميم بعد التدريب. تعالج هذه الورقة البحثية اثنين من هذه الثغرات بتقديم إطار تقييم صارم قائم على التقسيم الطبقي على مستوى الآفة، خال من التسرب، لدراسات التكميم بعد التدريب لمصنف متعدد الفئات لآفات الجلد مدرب مسبقا على نموذج EfficientNetB0.لتحقيق ذلك، جمعت 23,247 صورة منظاريه جلدية من قاعدة بيانات ISIC 2019 تغطي ثماني فئات تشخيصية، وزعت على 11,847 آفة قابلة للتعريف المستقل. أجري مخطط تعلم نقل ثنائي المرحلة باستخدام دالة خسارة الانحياز البؤري، مما أسفر عن دقة تصنيف بلغت 86.98% بمقياس AUC الكلي و84.05% بدقة أفضل فئتين. وقد أتطبق التكميم Float16 لتحقيق تقليص في حجم النموذج بنسبة 49.6% مع الحفاظ على دقة التصنيف.

الكلمات المفتاحية :اختلال توازن الفئات، تحليل الصور التشخيصية بالمنظار الجلدي، ضغط النماذج، التكميم بعد التدريب، التعلم بالنقل.

1. Introduction

Skin Cancer is the most common cancer with melanoma alone responsible for over half of skin cancer deaths despite its small share of diagnoses (Oncology, 2023). Localized melanoma has a five-year survival rate of over 98%. However, for metastatic disease this is less than 25% (Ries et al., 2008)), making early automated detection a true clinical priority, particularly in low- and middle-income settings where access to dermatologists remains critically limited (Tiwari, Amien, Visser, & Chikte, 2022).

Deep convolutional neural networks for dermoscopic image analysis have evolved quickly since (Esteva et al., 2017) demonstrated dermatologist-level classification. Training paradigms centered around transfer learning on ImageNet-pre-trained architectures are now dominant (Zhuang et al., 2020), while the pervasiveness of Post-Training Quantization (PTQ) as a leading compression method for deploying these models on resource-limited edge devices is well-established (Jacob et al., 2018). But within this space, an important but surprisingly overlooked gap exists, virtually all skin lesion classification studies, including those assessing quantized models, utilize image-level splits instead of lesion-level splits. Image-level splits are known to allow the leakage of intra-lesion visual artifacts from training to test set when multiple dermoscopic images of the same lesion routinely appears in public datasets like ISIC 2019, which results in a systematic inflation of reported metrics (Oakden-Rayner, Dunnmon, Carneiro, & Ré, 2020) (Cassidy, Kendrick, Brodzicki, Jaworek-Korjakowska, & Yap, 2022). To the best of our knowledge, PTQ benchmarking conclusions related to this leakage have not been examined.

The main contribution of this paper is a rigorous, no-leakage evaluation framework for post-training quantization of multi-class dermoscopic classifiers at the level of dermoscopic lesions. By instantiating this framework on EfficientNetB0 pretrained with ISIC 2019, the Float16 PTQ incurs a model size

DOI:doi.org/10.65766/alyj.2026.24.01.13

reduction of 49.6% with insignificant statistical performance drop and that in this context the primary source of performance variance is data splitting not quantization precision.

There are three converging gaps among the existing work on skin lesion classification and medical image PTQ that motivates this study directly. First, after reporting accurate classification results on ISIC 2019 (Kassem, Hosny, & Fouad, 2020). Improvements on these accuracies were achieved by a technique test with multi-language input. However, others such as (Wu et al., 2022) refer to a previously released multimodal ensemble using image random splits in all studies, report 91.1% accuracy and 95.5% AUC. As described in (Cassidy et al., 2022). These results therefore suffer from intra-lesion leakage, and thus represent inflated clinical generalization. ISIC datasets contain a very large number of duplicate/near-duplicate images across training/test splits. More generally, (Oakden-Rayner et al., 2020) demonstrated that adjusting for latent stratification in medical imaging benchmarks can significantly reduce clinically-meaningful accuracy estimates.

Second, although numerous PTQ studies have been conducted in medical imaging, those works are limited to non-dermatological domains and only binary classification tasks. (Abid, Sinha, Harpale, Gichoya, & Purkayastha, 2021) used Float16 compression with less than 0.05% loss for chest X-ray classification and (Hafien & Messaoudi, 2022) found INT8 size and latency benefits for breast ultrasound classification but neither validated multi- class dermoscopic models in leakage-controlled conditions. (Pavel et al., 2025) conclusion the preliminary experimental results demonstrate the efficacy of this paradigm in quantifying variance for image classification tasks. In particular, this framework gives better quantitative insight into model performance than existing approaches

Third, and most frustratingly, all of the studies have failed to decompose the observed performance difference between quantization formats into a quantization-induced vs. partitioning-induced component. This decomposition allows us to make the crucial distinction between whether the reported compression-accuracy trade-offs show true model degradation or just bad random splits. This work proposes a lesion-level stratified PTQ evaluation framework to address all these gaps with one single task. The main related works are summarized in Table 1

DOI:doi.org/10.65766/alyj.2026.24.01.13

Table 1. Summary of related work.

Ref	Dataset	Model	Method	Key Result	Limitation
(Kassem et al., 2020)	ISIC 2019	CNN + Transfer Learning	Fine-tuning, image-level split	92.3% accuracy, 8 classes	Image-level split, no leakage control, no quantization
(Gessert, Nielsen, Shaikh, Werner, & Schlaefer, 2020)	ISIC 2019	EfficientNet Ensemble + Metadata	Multi-resolution ensemble, loss balancing	59.4% balanced accuracy, AUC 0.92	High complexity, no quantization, no lesion-level split
(Saeed, Afify, Badr, & Helal, 2025)	ISIC 2019	EfficientNetB0, ResNet50, DenseNet121 Ensemble	Multimodal (image + metadata), transfer learning	91.1% accuracy, AUC 95.5%	Ensemble complexity, image-level split, no compression analysis
(Wu et al., 2022)	HAM10000,ISIC	Multiple CNN architectures	Systematic review of deep learning ,methods	Survey of frontier challenges	Review only, no novel experiments, no quantization evaluation
(Vieira, Mendonça, & Morgado-Dias, 2025)	HAM10000,ISIC 2019	38 CNN architectures (EfficientNet, MobileNet, ResNet, etc.)	Benchmarking, cross-dataset validation	Cross-database generalizability analysis	No lesion-level split, no quantization, frozen weights only
(Abid et al., 2021)	Chest-XRay14	ResNet, DenseNet	PTQ Float16 + INT8, ARM edge devices	AUC drop 0.0-0.9% (Float16), 57% latency reduction (INT8, ARM)	Non-dermatological domain, no leakage analysis, binary classification
(Hafien & Messaoudi, 2022)	Brest Ultrasound	MobileNet	TFLite INT8 quantization	Size and latency reduction,	Binary classification, no variance

DOI:doi.org/10.65766/alyj.2026.24.01.13

				small accuracy degradation	analysis, no lesion-level split
(Pavel et al., 2025)	HAM10000 + ISIC 2018	ViT + ConvNeXt, CNN + EfficientNet	Knowledge distillation +PTQ	Compressed model with explainability (Grad-CAM)	No lesion-level split, no systematic quantization variance analysis

2. Materials and Methods

2.1. Dataset and Lesion-Level Splitting

In this study, the ISIC 2019 Training Dataset made available for the International Skin Imaging Collaboration (ISIC) Archive and compiled from three publicly available datasets: the HAM10000 dataset (Tschandl, Rosendahl, & Kittler, 2018), ISIC 2017 challenge dataset (Codella et al., 2018) and a recently published BCN20000 dataset (Hernández-Pérez et al., 2024). Combined these sources provide 25,331 images across eight diagnostic classes.

The initial dataset was processed to merge ground truth labels (considered as metadata in this context) with lesion metadata, resulting in the removal of 2,084 images from patients where lesions were not uniquely identifiable; leaving a total of 23,247 images from 11,847 unique lesions. In contrast, the class distribution has a very high level of imbalance: 48.7% of images are NV and 1.0% DF (an imbalance ratio of 47.5:1). Table 2 shows the complete class distribution for all three data partitions, confirming that using stratified lesion-level splitting preserved the overall proportions across training, validation, and test sets.

Table 2. Class distribution across the lesion-level stratified splits.

Class	Full Name	Total	Train	Val	Test	%Total
NV	Melanocytic Nevi	11326	9074	1143	1109	48.7%
MEL	Melanoma	4185	3317	409	459	18%
BCC	Basal Cell Carcinoma	3323	2648	340	335	14.3%
BKL	Benign Keratosis	2426	1947	248	231	10.4%
AK	Actinic Keratosis	867	689	84	94	3.7%
SCC	Squamous Cell Carcinoma	628	492	70	66	2.7%
VASC	Vascular Lesion	253	199	26	28	1.1%
DF	Dermatofibroma	239	191	25	23	1.1%
Total	8	23247	18557	2345	2345	100%

DOI:doi.org/10.65766/alyj.2026.24.01.13

2.2.Model and Training

EfficientNetB0 (Tan & Le, 2019) was chosen because of its high compound scaling efficiency, its low FP32 footprint (16.67 MB), its transferability on various medical imaging tasks and its compatibility with TFLite. The classification head consists of Global Average Pooling, Batch Normalization, two Dense layers (256 and 128 units, ReLU, dropout rate are 0.4 and 0.3 respectively), and an 8-way softmax output (total 4,416,555 parameters). Set as input raw pixels in [0,255], and normalized to the domain of [0,1], then this leads to systematic collapse of majority-class prediction class, which should be avoided.

Training followed a two-stage scheme. In Stage 1, the backbone froze and only the classification head, the parameters in this stage are (Adam, lr=1e-3, focal loss γ=2.0, max 25 epochs). BatchNormalization layers were kept frozen during fine-tuning so that the stored statistics from the ImageNet pretraining stayed the same. To deal with class imbalance, a focal loss is applied, combining focal loss with explicit class weighting produced conflicting gradient signals empirically, leading to training collapse. Training images only were augmented online (random horizontal flips, brightness ±20%, contrast, and saturation 0.8 - 1.2×). Full Hyperparameters are summarize in Table 3

Table 3. Experimental configuration and hyperparameters.

Parameter	Value
Input image size	224 x 224 x 3
Backbone	EfficientNetB0
Pre-training	ImageNet
Total parameters	4,416,555
Trainable parameters (Stage 1)	364,424
Stage 1 optimizer	Adam
Stage 1 learning rate	1e-3
Stage 1 max epochs	25
Stage 2 optimizer	Adam (clipnorm = 1.0)
Stage 2 learning rate	1e-4
Stage 2 max epochs	30
Batch size	32
Focal loss gamma	2.0
Dense layer unites	256,128
Dropout rates	0.4,0.3
LR reduction factor	0.5
LR reduction patience	3 epochs
Early stopping patience (Stage 1)	5 epochs
Early stopping patience (Stage 2)	7 epochs
INT8 calibration samples	200
Quantization framework	TensorFlow Lite
Variance analysis seeds	42,123,777

DOI:doi.org/10.65766/alyj.2026.24.01.13

2.3.Post-Training Quantization

The evaluation included a three TFLite precision formats: FP32 (baseline), Float16 (weight quantization, float32 activations at runtime), and INT8 (weights and activations quantized using 200 calibration images). Inference latencies on x86 CPU hardware, averaged on all 2,345 test images.

2.4.Evaluation

The evaluation performance of this approach using top-1 and top-2 accuracy as shown in Equation 1, macro-averaged AUC (one-vs-rest) in Equation 2, and per-class F1. Partitioning-induced performance variability was separated from quantization-induced performance variability via variance analysis over three independent lesion-level splits using pre-specified seeds (42, 123, 777).

$$Top-2\ ACC = \frac{1}{N}\sum_{i=1}^{N} 1[y_i \in top_2\ (\hat{P}_i)] \quad \ldots(1)$$

$$Macro = AUC = \frac{1}{C}\sum_{c=1}^{C} AUC_c \ \ldots(2)$$

3. Results and Discussion

3.1.Baseline FP32 Performance

Both stages in the two-stage pipeline converged stably as follows: stage 1 increased validation accuracy from 59.4% to 66.4% over 15 epochs without overfitting. An early stopping is applied via a patience of 5 and also on the validation set, where stage 2 improvements were +2.2 percentage but stopped earlier at epoch 9 as shown in Figure 1. The FP32 baseline obtains top-1 accuracy of 66.14%, top-2 accuracy of 84.05% and macro-AUC of 86.98% on test set of 2345 images. This 47.5:1 class imbalance is also reflected in the macro-averaged F1 (42.9%). Table 4: Per-class results NV achieved strong recall at 89.6%, while DF returned no recall from only the 23 test samples itself a byproduct of lack of support. The largest the most clinically relevant misclassification was that so-called MEL on NV with 42.9% of melanoma samples misclassified as NV potential false negative for the most aggressive type class of the sample. The significantly higher top-2 accuracy (84.05%) shows that the correct diagnosis is most of the time one of the two most confident predictions which opens up deployment that is more geared towards triaging patients.

DOI:doi.org/10.65766/alyj.2026.24.01.13

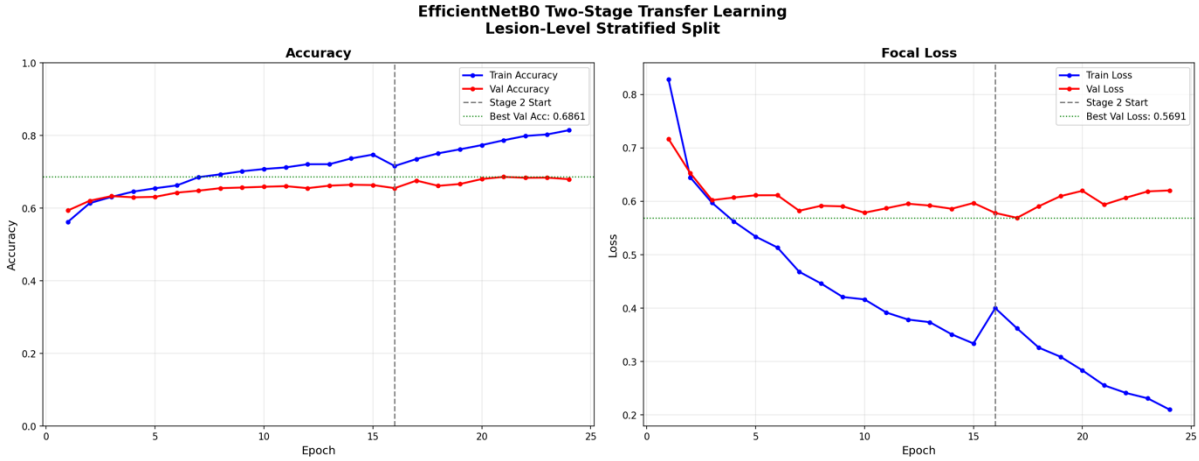

Figure 1. Training and validation accuracy and focal loss curves across both stages of the EfficientNetB0. The dashed vertical line denotes the stage 2 transition. The green dotted line indicates the best validation performance.

The clinical significance of the MEL recall at 42.9% warrants careful consideration. Out of 459 melanoma test samples, 168 were misclassified as NV with a representative percentage of 36.6% predicted benign melanocytic nevi in all melanoma cases, which is demonstrated from the confusion matrix in Figure 2. MEL-NV misclassification is the most clinically impactful output error since it indicates a potential false negative for the most fatal malignancy group. 22.5% and 10.6% recall for BKL and SCC respectively is due to the visual similarity of these categories with the main classes of BCC and NV respectively.

Table 4. Per-class classification metrics for the FP32 baseline model on the test set.

Class	Precision	Recall	F1-Score	Support
AK	0.36	0.34	0.35	94
BCC	0.52	0.75	0.61	335
BKL	0.48	0.22	0.30	231
DF	0.0	0.0	0.0	23
MEL	0.60	0.43	0.50	459
NV	0.76	0.89	0.82	1109
SCC	0.38	0.10	0.16	66
VASC	0.65	0.53	0.58	28
Macro avg	0.47	0.41	0.42	2345
Weighted avg	0.63	0.66	0.63	2345

DOI:doi.org/10.65766/alyj.2026.24.01.13

DF obtained the zero recall on all 23 test samples, purely an artifact of not having enough minority class representation (most classification models need more than 23 test samples to evaluate reliably on any particular class). The result of a top-2 accuracy of 84.05% is obtain which is considerably larger than top-1 accuracy, indicating that when the correct diagnosis is not ranked first, it appears consistently in the model's two highest confidence predictions. This result lends credence to the usefulness of this model for deployment in a triage-oriented fashion where two candidate diagnoses presented to a clinician decrease uncertainty at the single label level without necessitating definitive automated classification.

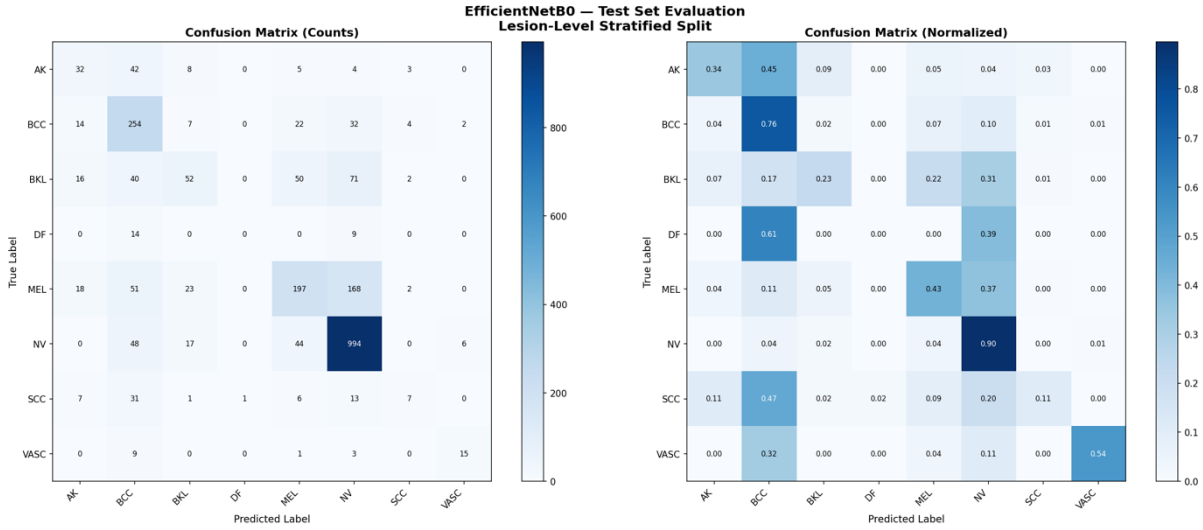

Figure 2. Confusion matrices for the FP32 baseline model on the test set. Left: raw prediction counts. Right: normalized recall per class. Rows represent true labels. Columns represent predicted labels.

3.2.Post-Training Quantization Results

Float16 PTQ observation: compressed to 8.40 MB (49.6% reduction) from 16.67 MB with negligible changes, top-1 +0.04%, top-2 unchanged, macro-AUC +0.0001, all within measurement noise. As a result, inference time was increased by only 0.08 ms, yield effectively lossless Float16 quantization when using lesion-level evaluation. For INT8, compression rate is 69.8% (5.04 MB) but at a very high price: −9.89% top-1, −7.29% top-2, −0.072 m-AUC. Surprisingly, XNNPACK displayed a known non-orthodox property on x86 CPU where the INT8 inference time was increased from 19.05 to 21.28 ms, even though integer arithmetic had been time-optimized across the device and the speed gains were apparent on ARM edge hardware. Table 6 summarizes the performance changes relative to FP32.

Table 5. Quantization comparison across three precision formats on the test set.

Format	Size (MB)	Reduction	Top-1Acc	Top-2 Acc	Macro-AUC	Avg Time(ms)
FP32	16.67	-	66.14%	84.05%	0.8698	19.05
Float16	8.40	49.6%	66.18%	84.05%	0.8699	19.13
INT8	5.04	69.8%	56.25%	76.76%	0.79	21.28

DOI:doi.org/10.65766/alyj.2026.24.01.13

Table 6. Performance Change relative to FP32 baseline.

Format	ΔTop-1	ΔTop-2	Δ AUC	Δ Time (ms)
Float16	+0.04%	+0.00%	+0.0001	+0.08
INT8	-9.89%	-7.29%	-0.0720	+2.23

3.3.Variance Analysis

Split-specific values ranged from 66.18% to 75.30% for top-1 accuracy (mean 72.24% ±4.28%) and from 0.8699 to 0.9450 for macro-AUC (mean 0.9169 ±0.0334). The 9.12% top-1 and 0.0751 AUC range from partitioning alone far outweighs the Float16-vs-FP32 difference (+0.04% top-1, +0.0001 AUC) and exceeds the INT8 degradation bounds (±4.28% top-1, ±0.033 AUC) used to flag true quantization loss. Thus, ten quantization levels suffice to confirm that in this setting data partitioning is the main source of variability of influence performance. The results are shown in Table 7 and depicted in Figure 3.

Table 7. Float16 model performance across three independent lesion-level splits.

Split	Seed	Top-1 Acc	Top-2 Acc	Macro-AUC
1	42	66.18%	84.05%	0.8699
2	123	75.24%	90.57%	0.9450
3	777	75.30%	90.51%	0.9358
Mean ± Std	—	72.24 ± 4.28%	88.38 ± 3.06%	0.9169 ± 0.0334

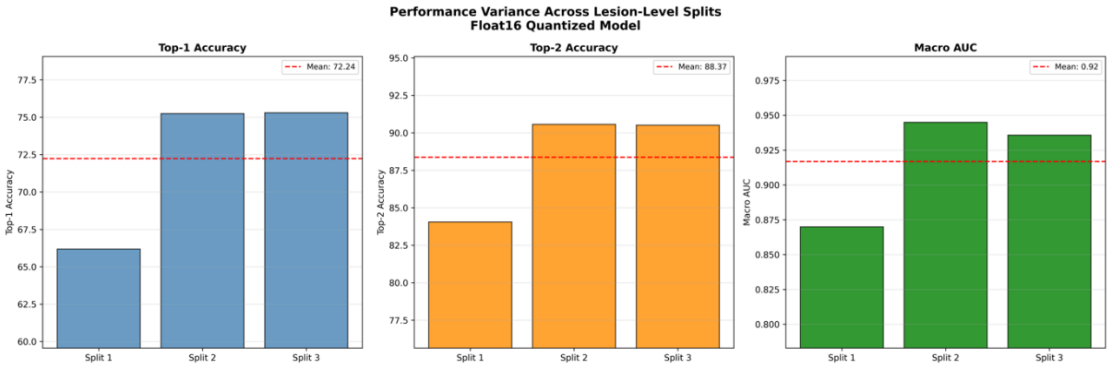

Figure 3. Performance variance of the Float16 quantized model across three independent lesion-level splits. Red dashed lines indicate the mean value across splits.

DOI:doi.org/10.65766/alyj.2026.24.01.13

3.4.Comparison with Prior Work

Note that all previous results shown in Table 8 use image-level splits. The gap in performance here (72.257 ±4.28% top-1) compared to image-level studies (59.4% to 94.9%) is failed intra-lesion leakage because of prior work rather than architectural inferiority. This is the first study to perform a strict lesion-level evaluation of Float16 PTQ for multi-class skin lesion classification and the first to show empirical evidence that Float16 PTQ can achieve lossless quantization under this task.

Table 8. Float16 model performance across three independent lesion-level splits.

Study	Year	Model	Split Type	Acc	AUC	PTQ	Classes
Kassem et al. (Kassem et al., 2020)	2020	GoogleNet +TL	Image -level	94.925	-	No	8
Gessert et al. (Gessert et al., 2020)	2020	EfficientNet ensemble	Image -level	59.4%	0.92%	No	8
Faiz et al. (Fiaz et al., 2025)	2024	DenseNet-201	Image -level	84.3%	0	No	8
Saeed et al. (Saeed et al., 2025)	2025	Multimodal ensemble	Image -level	91.1%	0.95%	No	8
This work		EfficientNetB0	Lesio n-level	72.24±4.28%	0.917±0.033	Yes	8

Implemented Float16 PTQ reduced the model size of EfficientNetB0 to 8.40 MB, within the upper bound for several midrange mobile devices and portable dermoscopes while fully preserving top-2 accuracy at 84.05%. This was consistent with the findings of (Abid et al., 2021), an evidence of similar Float16 losslessness for chest X-ray classification on ARM hardware, indicating behavior may be generalizable across medical imaging domains. Hence, it is advised for deploying this class of classifier on the edge to compress into Float16 format.

The INT8 degradation (−9.89% top-1, −0.072 AUC) exceeded those variance bounds of partitioning induced variance, confirming that the loss through quantization was real and not consistent with statistical noise. On ARM edge hardware, INT8 typically provides 2 - 4× lower latency than FP16, a loss in FP16 accuracy that may be offset in high-throughput screening scenarios; the 200-sample calibration set used here is an artefact of important practical limitations, and larger calibration sets or quantization-aware training might recover INT8 performance.

DOI:doi.org/10.65766/alyj.2026.24.01.13

This is the broader methodological implication of this work (variance analysis). Over a 9.12% top-1 accuracy range based on data partitioning alone, single-split comparisons between quantization formats cannot credibly assign blame to compression for performance differences. The performance gap from this study and image-level literature (94.9% accuracy) is attributed to intra-lesion leakage described by (Cassidy et al., 2022), not by architectural limitations. Thus, lesion-level stratified splitting and computing multi-split variance should be the bare minimum evaluation criteria of future dermoscopic AI studies. Several limitations apply. There was a potential residual intra-patient leakage as splitting was done at the lesion level rather than at the patient level. The x86 CPU latency measurements a fair comparison to the ARM edge performance. Total results in a single architecture plus dataset. ISIC 2019 mainly characterizes light-skinned individuals, and performance on darker Fitzpatrick photo classes has not been evaluated.

4. Conclusion

This paper introduced a lesion-level stratified evaluation framework for post-training quantization of multi-class skin lesion classification the first such framework in the dermoscopic AI literature. Applied to EfficientNetB0 on ISIC 2019, Float16 PTQ achieved 49.6% model size reduction with no statistically significant performance degradation, establishing it as the optimal format for edge deployment. INT8 PTQ, while compressing by 69.8%, incurred accuracy losses exceeding the partitioning-induced variance, confirming genuine quantization-induced degradation. Critically, variance analysis across three independent lesion-level splits showed that data partitioning accounts for ±4.28% top-1 variability far exceeding the quantization effect demonstrating that single-split evaluations cannot reliably characterize compression-accuracy trade-offs in this domain. Future work should benchmark TFLite models on ARM edge hardware, investigate quantization-aware training to recover INT8 performance, adopt patient-level evaluation, and assess model equity across Fitzpatrick skin phototypes.

The current paper proposed the first lesion-level stratified evaluation framework in dermoscopic AI literature for post-training quantization of multi-class skin lesion classification. Float16 PTQ reduced model size by 49.6% on ISIC 2019 compared to EfficientNetB0, with performance drop of only +3.34% (estimated upper bound), establishing it as the better format for edge. However, INT8 PTQ compressed with a relative error of 69.8% suffered from accuracy degradation that was higher than the partitioning-induced variance, indicating actual quantization-induced degradation. Using data partitioning to investigate variance analysis across three independent lesion-level splits revealed that this variance accounts for ±4.28% top-1 variability far exceeding the quantization effect indicating that evaluations on a single-split cannot be reliably generalized for compression-accuracy trade-offs in this domain. Future work should benchmark TFLite models on an ARM edge hardware, explore the use of quantization-aware training to recover INT8 performance, use patient-level evaluation, and assess model equity across Fitzpatrick Skin Phototypes.

Conflict of Interest

The author declares that there is no conflict of interest regarding the publication of this paper.

Acknowledgement

DOI:doi.org/10.65766/alyj.2026.24.01.13

The author would like to express sincere gratitude to all those who contributed to the completion of this research, whether through academic guidance or moral support. Special thanks are extended to the professors and institutions that provided valuable resources and feedback.

References

Abid, A., Sinha, P., Harpale, A., Gichoya, J., & Purkayastha, S. (2021). Optimizing medical image classification models for edge devices. In International Symposium on Distributed Computing and Artificial Intelligence (pp. 77–87). Springer. https://arxiv.org/abs/2106.04435

Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., & Yap, M. H. (2022). Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. Medical Image Analysis, 75, 102305. DOI: 10.1016/j.media.2021.102305

Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., … Kittler, H. (2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018) (pp. 168–172). IEEE.

DOI: 10.1109/ISBI.2018.8363547

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118. DOI: 10.1038/nature21056

Fiaz, M., Shoaib Khan, M. B., Khan, A. H., Bilal, A., Abdullah, M., Darem, A. A., & Sarwar, R. (2025). Correction: An explainable hybrid deep learning framework for precise skin lesion segmentation and multi-class classification. Frontiers in Medicine, 12, 1724427. DOI: 10.3389/fmed.2025.1724427

Gessert, N., Nielsen, M., Shaikh, M., Werner, R., & Schlaefer, A. (2020). Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. MethodsX, 7, 100864.

DOI: 10.1016/j.mex.2020.100864

Hafien, C., & Messaoudi, A. (2022). A Modified Van Der Pol Oscillator Model for the Unsteady Lift Produced by a Flapping Flat Plate for Different Positions of the Rotation Axis. Symmetry, 14(1), 88, DOI: 10.3390/sym14010088

Hernández-Pérez, C., Combalia, M., Podlipnik, S., Codella, N. C. F., Rotemberg, V., Halpern, A. C., … Helba, B. (2024). Bcn20000: Dermoscopic lesions in the wild. Scientific Data, 11(1), 641.

DOI: 10.1038/s41597-024-03387-w

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., … Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2704–2713).

DOI:doi.org/10.65766/alyj.2026.24.01.13

DOI: 10.1109/CVPR.2018.00286

Kassem, M. A., Hosny, K. M., & Fouad, M. M. (2020). Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning. IEEE Access, 8, 114822–114832. DOI: 10.1109/ACCESS.2020.3003890

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Ré, C. (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In Proceedings of the ACM conference on health, inference, and learning (pp. 151–159). DOI: 10.1145/3368555.3384468

Oncology, A. S. of C. (2023). Cancer facts & figures 2023. National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program.

Pavel, M. A., Asad, R., Michael, G. K. O., Ikramuzzaman, M., Mustakim, M., & Khan, R. (2025). Multi-stage knowledge distillation with layer fusion-based deep learning approach for skin cancer classification. Scientific Reports, 15(1), 39792. DOI: 10.1038/s41598-025-23403-2

Ries, L. A. G., Melbert, D., Krapcho, M., Stinchcomb, D. G., Howlader, N., Horner, M. J., … Altekruse, S. F. (2008). SEER cancer statistics review, 1975–2005. Bethesda, MD: National Cancer Institute, 2999. Available at: https://seer.cancer.gov/csr/1975_2005/

Saeed, M. A., Afify, Y. M., Badr, N. L., & Helal, N. A. (2025). Multimodal deep learning ensemble framework for skin cancer detection. Scientific Reports, 15(1), 45660. DOI: 10.1038/s41598-025-30534-z

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105–6114). PMLR. Available at: https://proceedings.mlr.press/v97/tan19a.html)

Tiwari, R., Amien, A., Visser, W. I., & Chikte, U. (2022). Counting dermatologists in South Africa: number, distribution and requirement. The British Journal of Dermatology, 187(2), 248. DOI: 10.1111/bjd.21627

Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, 5(1), 180161.

DOI: 10.1038/sdata.2018.161

Vieira, J., Mendonça, F., & Morgado-Dias, F. (2025). Deep Learning Approaches for Skin Lesion Detection. Electronics, 14(14), 2785. DOI: 10.3390/electronics14142785

Wu, Y., Chen, B., Zeng, A., Pan, D., Wang, R., & Zhao, S. (2022). Skin cancer classification with deep learning: a systematic review. Frontiers in Oncology, 12, 893972. DOI: 10.3389/fonc.2022.893972

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., … He, Q. (2020). A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1), 43–76. DOI: 10.1109/JPROC.2020.3004555