AyahVerse at NakbaArchiveClassifier Shared Task: Architectural Trade-offs and Decision Calibration for Humanitarian Image Classification

Ibad-ur-Rehman Rashid, Akhtar Ali
Government Post Graduate College, Mansehra, Affiliated with Hazara University, Pakistan;
Northeastern University, China
ibad@gcm.edu.pk, ata@mails.neu.edu.cn

Abstract

This paper presents our submission to the Nakba-NLP 2026 Shared Task on binary image classification, where the goal is to categorize images of Gaza infrastructure as destroyed or intact. To address the challenges of class imbalance and resource-constrained deployment, we evaluated three convolutional architectures: ResNet50, MobileNetV2, and EfficientNet-B0, combined with a post-hoc threshold optimization step. Our results show that lightweight architectures are competitive with heavier models for this task, with EfficientNet-B0 achieving the highest Test F1-score of 0.85 despite having significantly fewer parameters than ResNet50. We further investigated the effect of input resolution, finding that increasing resolution improved ResNet50's performance, though it remained below lightweight alternatives. Finally, we demonstrate that shifting the binary decision threshold from the default 0.50 to an optimized 0.45 improved ResNet50's Test F1 from 0.79 to 0.81 by recovering recall for the minority destroyed class. Notably, this adjustment was only needed for ResNet50, while EfficientNet-B0 and MobileNetV2 performed best at the default 0.50, suggesting that larger models are more prone to majority-class bias. Overall, these results provide a systematic analysis of architectural efficiency and threshold behavior under class imbalance, offering practical insights for damage classification in resource-constrained crisis settings.

Keywords: Image Classification, EfficientNet, MobileNet, Edge Computing, Threshold Optimization, Humanitarian AI

1. Introduction

The Nakba Image Classification task(Abrahams et al., 2026) highlights the urgent need to categorize millions of images archived from social media by distinguishing destroyed infrastructure from intact environments. While academic competitions typically prioritize absolute accuracy through massive ensembles, practical deployment in humanitarian contexts, such as field hospitals or local NGO servers, often requires models that can run without high-end GPUs.

This paper presents the AyahVerse team's binary classification system. Instead of solely maximizing predictive metrics via scaling, we evaluate the trade-offs necessary for field-ready deployment. We structure our experiments around three simple questions:

1. How do lighter, mobile-friendly models (like MobileNetV2 and EfficientNet-B0) compare to standard, heavy models (like ResNet50) for this specific task?
2. Does using larger, higher-resolution images actually help the model spot damage, or does it just confuse the model with noisy details?
3. How much does adjusting the standard decision threshold help the model handle datasets where one class (intact buildings) heavily outnumbers the other (destroyed buildings)?

Our code is available at Github.Our code is available at Github.[1]

2. Background

2.1. Related Work

Our system design is informed by three intersecting areas of research: humanitarian image classification, efficient deep learning architectures, and decision calibration for imbalanced data.

Crisis Informatics and Image Classification

The automated analysis of social media imagery during crises has gained significant traction. Datasets like CrisisMMD (Alam et al., 2018) have supported research into categorizing disaster-related images to aid humanitarian response. However, much of the existing literature focuses on maximizing absolute accuracy using massive ensemble models (Nguyen et al., 2017). Our work pivots from this trend by prioritizing deployability in resource-constrained humanitarian contexts, specifically for the Nakba-NLP damage assessment task.

[1] https://github.com/Ebad-urRehman/nakba_archive_efficient_classification

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 212–216

Efficient Architectures for Edge Deployment

While deep residual networks like ResNet (He et al., 2015) have historically dominated image classification, their computational overhead makes them unsuitable for edge deployment (e.g., field laptops or localized NGO servers). Recent advancements have focused on mobile-optimized architectures. MobileNetV2 (Sandler et al., 2019) introduced inverted residuals and linear bottlenecks to significantly reduce parameter counts. Similarly, EfficientNet (Tan and Le, 2020) proposed a compound scaling method that balances network depth, width, and resolution. We directly compare these architectures to determine the optimal trade-off between parameter count and damage-recognition capability.

Mitigating Class Imbalance

This dataset is highly imbalanced, as intact infrastructure heavily outnumbers destroyed infrastructure. Standard approaches to class imbalance include resampling techniques or cost-sensitive learning, such as Focal Loss (Lin et al., 2018), which reduces the loss contribution of easy, well-classified examples to focus training on hard cases. Beyond the loss function, shifting the binary decision threshold (threshold moving) is a proven technique to calibrate classifiers against asymmetric misclassification costs (Provost et al., 1998). We extend this by searching for the optimal threshold during the validation phase to improve minority-class recall without severely degrading precision.

2.2. Dataset

The shared task requires binary classification on a dataset of 2,001 Instagram images collected between October 2023 and December 2025. The data features a significant class imbalance: 706 images (35.3%) of direct conflict and destruction versus 1,295 (64.7%) intact scenes.

Class	Number of Images	Percentage
Destroyed Infrastructure	706	35.3%
Intact Infrastructure	1,295	64.7%
Total	**2,001**	**100%**

Table 1: Class distribution in the Nakba Image Classification dataset showing significant imbalance (35.3% destroyed vs. 64.7% intact).

3. System Overview

To mitigate the class imbalance inherent to the dataset, we adopt two complementary strategies: replacing standard binary cross-entropy with a customized Focal Loss ($\alpha = 1$, $\gamma = 2$) at training time, and applying post-hoc threshold optimization at inference time. Together, these address both the learning and prediction stages of the classification pipeline.

3.1. Data Preprocessing and Augmentation

All images were normalized using ImageNet mean and standard deviation values. During training, we applied random resized crops, horizontal flips, random rotation, and color jittering to improve generalization. All models were evaluated at a standard resolution of 224×224, with an additional variant ResNet50 tested at 334×334 to investigate the effect of higher resolution.

3.2. Architectural Pipelines

We established three distinct pre-trained pipelines to evaluate architectural trade-offs:

- **Heavyweight Baseline (ResNet50):** A 50-layer deep residual network ($\sim$25.5M parameters).
- **Lightweight Competitor (MobileNetV2):** A mobile-optimized architecture ($\sim$3.4M parameters).
- **Balanced Competitor (EfficientNet-B0):** A compound-scaling architecture ($\sim$5.3M parameters).

3.3. Decision Calibration

Standard binary classifiers apply a default decision boundary of $t = 0.5$ to the sigmoid probability output $p_i = P(\hat{y}_i = 1|x_i)$. However, in highly imbalanced disaster datasets, this static boundary often suppresses minority-class predictions. To account for the asymmetric costs of misclassifying destroyed infrastructure, we parameterize the predicted class label $\hat{y}_i(t)$ as a function of an adjustable threshold t:

$$\hat{y}_i(t) = \begin{cases} 1, & \text{if } p_i \geq t \\ 0, & \text{otherwise} \end{cases}$$

Rather than relying on the default, we implement a post-hoc calibration step. During the validation phase, we perform a grid search over a bounded interval $t \in [0.3, 0.75]$ to find the optimal threshold t^* that maximizes the Macro F1 score on the validation set (V):

$$t^* = \underset{t \in [0.3, 0.75]}{\arg\max} \left(\text{Macro F1}(y_V, \hat{y}(t)_V)\right)$$

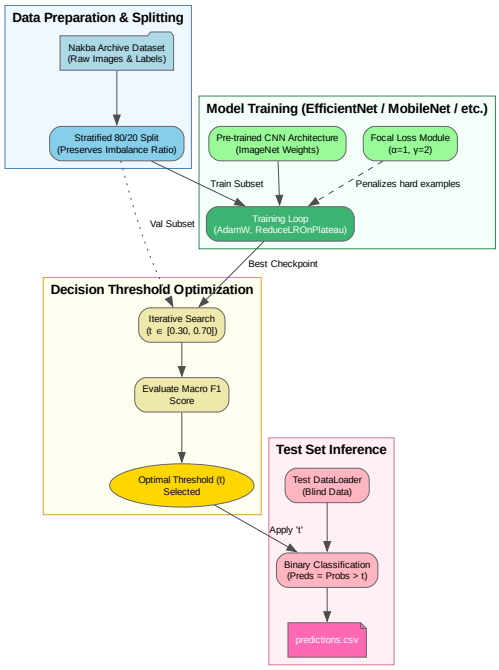

Figure 1: System architecture highlighting data splitting, EfficientNet-B0 training, and post-hoc threshold optimization.

While implemented as a lightweight grid search over $t \in [0.3, 0.75]$, this step is modular and can be applied to any binary classifier independently of the training procedure. The threshold yielding the highest validation metric is then frozen as the operational boundary for all final test set inferences as shown in Figure 3.

4. Experimental Setup

We combined the official training and validation sets provided by the shared task, which makes a total of 1,599 images, and performed an 80/20 stratified split to create our own training (1,279 images) and validation (320 images) sets for all experiments. All models were trained on a single T4 GPU using the AdamW optimizer (initial lr=1×10^{-4}), a ReduceLROnPlateau scheduler (patience=3), and the best checkpoint was saved based on validation Macro F1. All three architectures were evaluated under identical hyperparameter configurations, with only the model swapped between runs. Standard models processed 224×224 RGB imagery, while a secondary ResNet50 variation was scaled to 334×334 to test high-resolution sensitivity.

Dataset Split	Total Images
Combined (Train + Val)	1,599
Training Split (80%)	1,279
Validation Split (20%)	320

Table 2: Combined official Train+Val sets (1,599 images) with 80/20 stratified train-validation split for experiments.

5. Results and Analysis

Our quantitative evaluation revealed distinct performance dynamics across architectures and thresholds. Table 3 outlines the metric performance on the blind test set, where our best official submission achieved a Test Macro F1 of 0.85 using EfficientNet-B0.

Model Architecture	Resolution	Threshold (t)	Val Macro F1	Test Macro F1
ResNet50 (Baseline)	224x224	0.50	0.81	0.79
ResNet50 (High Res)	334x334	0.50	0.87	0.83
MobileNetV2	224x224	0.50	0.83	0.84
EfficientNet-B0	**224x224**	**0.50**	**0.89**	**0.85**
ResNet50 (Optimized)	224x224	0.45	0.82	0.81

Table 3: Test set evaluation across architectural, resolution, and threshold variations.

5.1. Architectural and Resolution Trade-offs

EfficientNet-B0 achieved the strongest performance in this task, achieving an F1-score of 0.85 while remaining significantly lighter than ResNet50. Surprisingly, MobileNetV2 (0.84) also outperformed the heavier ResNet50 baseline (0.79). This indicates that massive parameter depth may not be strictly necessary for identifying general disaster markers, as depthwise separable convolutions are capable of capturing structural damage patterns effectively.

Furthermore, increasing spatial resolution to 334×334 improved ResNet50's test performance to 0.83, though it still fell below MobileNetV2 (0.84) and EfficientNet-B0 (0.85) as shown in Table 3.

5.2. Threshold Optimization Impact

During validation, the standard $t = 0.5$ yielded an F1 of 0.81. Shifting the threshold to $t = 0.45$ yielded a marginal validation improvement (0.81 $\rightarrow$ 0.82), which nonetheless translated to a more meaningful gain on the blind test set (0.79 $\rightarrow$ 0.81). Notably, this adjustment was only necessary for ResNet50, as EfficientNet-B0 and MobileNetV2 achieved their best results at the default $t = 0.50$ despite the same class imbalance, suggesting that larger models are more susceptible to majority-class bias.

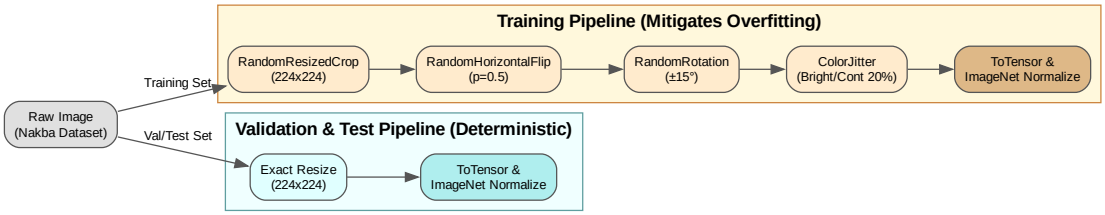

Figure 2: Data preparation and augmentation pipelines. The training set undergoes heavy spatial and color transformations to mitigate overfitting, while the validation and test sets follow a deterministic resizing and normalization path.

Threshold Optimization on Validation Set

Threshold Sweep Results	
Threshold (t)	**Macro F1**
0.35	0.7977
0.40	0.8039
0.45	**0.8214**
0.50	0.8197
0.55	0.7898
...	...

Default (0.5): 0.8197 F1

Optimal (0.45): 0.8214 F1

Improvement: +0.0017

Figure 3: Threshold optimized on validation set (t=0.45) and applied to test set, improving ResNet50 F1 from 0.79 to 0.81.

6. Conclusion

Our experiments demonstrate that lightweight architectures are not only computationally efficient but also more robust to class imbalance for this task. EfficientNet-B0 achieved the highest Test F1 of 0.85 while requiring far fewer parameters than ResNet50. Increasing ResNet50's input resolution improved test performance (0.79 $\rightarrow$ 0.83), though it remained below lightweight alternatives. Most notably, threshold calibration was only necessary for ResNet50, while EfficientNet-B0 and MobileNetV2 performed well at the default threshold, suggesting that larger models may require additional post-hoc calibration in imbalanced settings. These findings contribute a systematic comparison of architectural efficiency and threshold behavior under class imbalance, offering practical insights for future crisis classification systems deployed in resource-constrained environments.

7. Bibliographical References

Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. The nakbaarchiveclassifier shared task on nakba image classification. In Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026), Palma, Mallorca, Spain.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natural disasters.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection.

Dat Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. 2017. Robust classification of crisis-related data on social networks using convolutional neural networks. Proceedings of the International AAAI Conference on Web and Social Media, 11(1):632–635.

Foster J. Provost, Tom Fawcett, and Ron Kohavi. 1998. The case against accuracy estimation for comparing induction algorithms. In Proceedings of the Fifteenth International Conference on Machine Learning, ICML '98, page 445–453, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2019.

Mobilenetv2: Inverted residuals and linear bottlenecks.

Mingxing Tan and Quoc V. Le. 2020. Efficientnet: Rethinking model scaling for convolutional neural networks.