ZAHIRA BOULANOUAR at NakbaArchiveClassifier Shared Task: Detecting Infrastructure Destruction in Gaza with a ConvNeXt Ensemble

Zahira Boulanouar

Euromed University of Fez
zahira.boulanouar@eidia.ueuromed.org

Abstract

We present our third-place submission to the Nakba Image Classification Shared Task at LREC-COLING 2026, which requires binary classification of Instagram images from Gaza into destruction (damaged or destroyed infrastructure) versus not_destruction. Our system fine-tunes a ConvNeXt-Tiny backbone within a five-fold stratified cross-validation framework, combining Focal Loss, weighted random sampling, exponential moving average (EMA) weight stabilization, test-time augmentation (TTA), and out-of-fold (OOF) decision threshold calibration. Our system achieves an official test macro F1 of 0.8893 and 90.05% accuracy, placing third among all participants and within 0.02 F1 of the winning system (0.91), demonstrating that a 28M-parameter convolutional architecture with principled training strategies is highly competitive with much larger models.

Keywords: image classification, infrastructure damage detection, ConvNeXt, ensemble learning, Focal Loss, humanitarian AI, Gaza, Nakba

1. Introduction

Since October 7, 2023, Palestinian journalists and content creators in Gaza have extensively documented the ongoing conflict through social media. The SaltPillar project and the NakbaArchiveClassifier shared task (Abrahams et al., 2026) have archived and organized millions of Instagram images from this period. Automatically categorizing this archive is essential for enabling structured access for legal documentation, humanitarian reporting, and academic research.

The Nakba Image Classification Shared Task frames this as binary image classification: destruction (bombed buildings, rubble, damaged roads) versus not_destruction (intact infrastructure, people, unrelated content), evaluated by macro-averaged F1. Key challenges include moderate class imbalance (35% vs. 65%), high intra-class visual variation, and a small training set (1,400 images).

We address these with a ConvNeXt-Tiny (Liu et al., 2022) ensemble integrating Focal Loss (Lin et al., 2017), EMA weight averaging, weighted sampling, horizontal-flip TTA, and OOF threshold calibration. Our system ranked 3rd out of all participants (test macro F1: 0.8893; 1st place: 0.91), outperforming a larger EfficientNetV2-M (Tan and Le, 2021) single-model baseline by 10.3 F1 points confirming that ensemble strategies and calibration matter more than raw model capacity.

2. Related Work

Automated damage assessment from imagery has been studied primarily using satellite data (Gupta et al., 2019). Social media and ground-level imagery from active conflict zones introduce additional challenges: variable viewpoints, compression artifacts, and significant distributional shift. The Nakba dataset is, to our knowledge, the first large-scale labeled collection of conflict-zone infrastructure damage imagery.

ConvNeXt (Liu et al., 2022) modernizes the ResNet architecture with larger kernels, depthwise convolutions, GELU activations, and inverted bottlenecks achieving state-of-the-art accuracy while remaining efficient and well-suited to fine-tuning on small datasets. Deep ensemble methods (Lakshminarayanan et al., 2017) reduce prediction variance and improve calibration without architectural changes, while Focal Loss (Lin et al., 2017) addresses class imbalance by down-weighting well-classified examples.

3. Data

The dataset, released as part of the NakbaArchiveClassifier shared task (Abrahams et al., 2026), provides 2,001 labeled Instagram images (Palestinian journalists and content creators, Gaza, Oct. 2023–Dec. 2025). We use the official 1,400/199 train/validation split. The class distribution is moderately imbalanced: 64.7% not_destruction (1,295 images) vs. 35.3% destruction (706 images), maintained via stratification across all splits. The held-out test set comprises 500 images released February 10–17, 2026.

4. System Description

4.1 Architecture Overview

Figure 1 illustrates our end-to-end pipeline. At its core, the pipeline processes SaltPillar Instagram images from Gaza through a two-phase workflow. In the training phase, a weighted random sampler constructs balanced mini-batches, and ConvNeXt-Tiny models are fine-tuned with Focal Loss (α=0.35, γ=2.0), EMA weight stabilization, and stratified k-fold cross-validation. During inference, test images undergo horizontal-flip test-time augmentation (TTA), ensemble

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 290–293

averaging across folds, and out-of-fold (OOF) threshold calibration at τ=0.41 to produce destruction/not_destruction predictions.

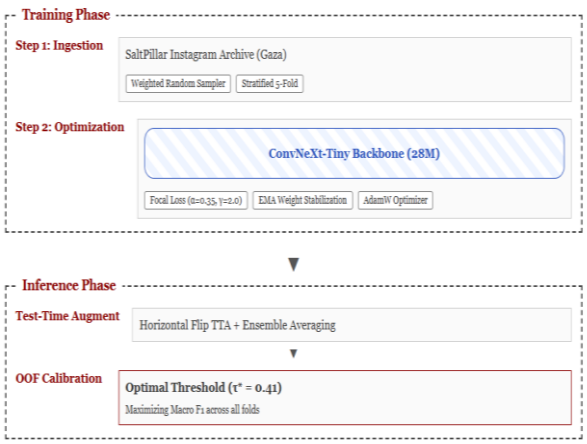

Figure 1: End-to-end system pipeline.

4.2 ConvNeXt Architecture

ConvNeXt (Zhuang Liu et al., 2022) is a modernized convolutional neural network architecture designed to close the performance gap between traditional CNNs and transformer-based vision models such as Vision Transformer. Instead of replacing convolutions with attention mechanisms, ConvNeXt revisits the standard ResNet design and incorporates several architectural improvements inspired by transformer models.

ConvNeXt blocks rely on depthwise separable convolutions, large kernel sizes (7×7), layer normalization, and GELU activations. These design changes increase the receptive field and improve gradient flow while maintaining the efficiency advantages of convolutional networks.

The ConvNeXt-Tiny variant used in this work contains approximately 28 million parameters and follows a hierarchical architecture composed of four stages with progressively decreasing spatial resolution and increasing channel depth. This design preserves the multi-scale feature extraction capability typical of convolutional architectures while benefiting from modern training techniques.

4.3 Training Configuration

Hyperparameter	Value
Input resolution	384 × 384
Optimizer	AdamW, lr = 1.2×10^{-4}, wd = 0.02
LR schedule	Cosine annealing
Max epochs	20 (early stop, patience = 6)
Loss	Focal Loss (α=0.35, γ=2.0)
EMA decay	0.999
Cross-validation	5-fold stratified

Table 1: Hyperparameters applied uniformly across all five folds.

Focal Loss re-weights gradients to focus training on hard examples, where p_t is the predicted probability for the ground-truth class, α_t=0.35 balances the minority destruction class, and γ=2.0 suppresses well classified negatives.

$$FL(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t)$$

Table 1 summarizes the hyperparameters applied uniformly across all five folds.

Weighted Random Sampling assigns each sample a weight inversely proportional to its class frequency per fold, ensuring approximately balanced mini-batches and complementing Focal Loss at the batch-construction level.

EMA maintains a shadow copy of model parameters (decay = 0.999) updated throughout training. EMA weights are used for all evaluation and saved as fold checkpoints, providing implicit within-run model averaging.

Training augmentation: random horizontal flip (p=0.5), color jitter (brightness, contrast, saturation ±0.2; hue ±0.05; p=0.6), random rotation (±10°). Validation and test: resize and normalize only.

4.4 Ensemble, TTA, and Threshold Calibration

Five models are trained on independent folds. At inference, TTA averages predictions on the original image and its horizontal flip per fold. The ensemble probability is the mean across all five folds.

$$p_{ens}(x) = \frac{1}{K}\sum_{i=1}^{K} p_i(x)$$

The default threshold of 0.5 is suboptimal for imbalanced macro F1. We sweep τ ∈ [0.10, 0.70] at step 0.004 and select the value maximizing macro F1 on OOF predictions, yielding τ* = 0.41, applied directly to validation and test predictions.

5. Results

5.1 Competition Standing

The shared task evaluation metric is macro-averaged F1, which balances precision and recall across both classes.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

$$\text{Macro F1} = \frac{1}{N} \sum_{i=1}^{N} F1_i$$

Our system ranked third among all participants, within 0.02 macro F1 of the top entry.

Table 2 shows the partial leaderboard.

Rank	Team	Macro F1	Acc.
1	HCMUS_TheFangs	0.8991	90.80%
2	rahaf_jaber	0.8952	90.55%
3	**zahira_blr (Ours)**	**0.8893**	**90.05%**

Table 2: Official leaderboard (all 3 participants)

5.2 Cross-Validation and Overall Performance

Per-fold macro F1 ranges from 0.8916 to 0.9229 (mean: 0.9078), indicating stable performance. Folds 2 and 5 triggered early stopping. Table 3 summarizes all evaluation splits.

Split	Macro F1	Accuracy	τ
OOF (train)	0.8975	90.86%	0.41
Validation	0.8515	86.93%	0.41
Test (official)	0.8893	90.05%	0.41

Table 3: Performance across all evaluation splits (τ*=0.41).

The OOF report (Table 4) shows well-balanced per-class performance (FN = 88, FP = 40). The test score (0.8893) exceeding validation (0.8515) suggests the 199-image validation set is noisier or slightly harder than the 500-image test set.

Class	P	R	F1	N
not_destruction	0.908	0.956	0.931	906
destruction	0.910	0.822	0.864	494
Macro avg	**0.909**	**0.889**	**0.898**	**1400**

Table 4: OOF classification report with τ*=0.41.

5.3 Comparison with EfficientNetV2-M

Table 5 compares our ensemble against a single EfficientNetV2-M trained with the same Focal Loss and augmentation setup, without cross-validation ensembling. Despite having nearly twice the parameters and higher destruction recall (87.1%), EfficientNetV2-M produces 32 false positives that collapse the not_destruction F1, falling 10.3 points below our ensemble. This confirms that ensembling and calibration outweigh raw model capacity.

Model	Val F1	Acc.
EfficientNetV2-M (single)	0.7485	79.4%
ConvNeXt-Tiny (ens.)	**0.8515**	**86.9%**

Table 5: Validation comparison. EfficientNetV2-M not submitted officially.

6. Discussion and Conclusion

Five design decisions drove our performance: (1) the pretrained ConvNeXt-Tiny backbone for low-data transfer; (2) Focal Loss + weighted sampling for joint imbalance correction; (3) EMA for within-run weight stabilization; (4) five-fold ensemble averaging for variance reduction; and (5) OOF threshold calibration (τ*=0.41) to balance class-wise F1.

The 0.02 gap to first place could be narrowed by Mixup/CutMix augmentation, multi-scale TTA, stochastic weight averaging, or multimodal fusion with caption text and metadata. Our models operate on pixels alone and do not exploit auxiliary signals that could disambiguate ambiguous cases (partially damaged buildings, construction sites).

The system should be treated as a decision-support tool not an autonomous classifier given the evidentiary and humanitarian sensitivity of the data. We presented a ConvNeXt-Tiny ensemble system that ranked 3rd in the Nakba Image Classification Shared Task with a test macro F1 of 0.8893, within 0.02 of the top entry and 10.3 points above a larger single-model baseline. We release this pipeline as a reproducible baseline for humanitarian image classification research.

7. Acknowledgments

We thank the Nakba-NLP 2026 Workshop organizers and the Palestinian journalists and content creators whose documentation makes this work possible.

8. Bibliographical References

Abrahams, Alexei, 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, May.

Dosovitskiy, A. et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Gupta, R. et al. (2019). xBD: A dataset for assessing building damage from satellite imagery. In *CVPR Workshops*.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986.

Tan, M. and Le, Q. V. (2021). EfficientNetV2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 10096–10106.

Wightman, R. (2019). PyTorch Image Models (timm). https://github.com/rwightman/pytorch-image-models.