Pixel at NakbaArchiveClassifier Shared Task: ConvNeXt-Based Ensemble for Destruction Detection

Rahaf Jaber
Independent Researcher
Al-Ahsa, Saudi Arabia
rahaf.fathi.jaber@gmail.com

Abstract

This paper describes our submission to the Nakba Image Classification Shared Task at the Nakba-NLP 2026 workshop. The task requires binary classification of social media images into two categories: *destruction* and *not_destruction*. The dataset includes approximately 1,600 annotated development images and 400 held-out test images, collected from Instagram posts published in Gaza between October 2023 and December 2025. High variability in viewpoint, lighting, and image quality, coupled with the inherent complexities of identifying structural damage in dense urban environments, makes this task particularly challenging. Our system utilizes a pretrained ConvNeXt-Tiny backbone fine-tuned through a stratified 5-fold cross-validation framework. To mitigate class imbalance, we implement a weighted cross-entropy loss function. During the inference phase, we employ an ensemble strategy that averages predictions across all five fold-specific models, and test-time augmentation (TTA) is applied to enhance robustness. The final ensemble achieved a Macro F1-score of 0.8952 and an accuracy of 0.9055 on the official test set. Our results suggest that the integration of modern convolutional architectures with robust ensembling and augmentation strategies provides a reliable baseline for automated destruction detection.

Keywords: image classification, destruction detection, ensemble learning, transfer learning, cross-validation

1. Introduction

The Nakba Image Classification Shared Task, organized as part of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), addresses the critical problem of automatically distinguishing images depicting destroyed buildings from those showing intact structures (Abrahams et al., 2026). As large-scale digital archives of conflict-zone imagery grow, manual categorization becomes increasingly infeasible. Automatic classification is therefore essential for improving the accessibility of these archives, supporting humanitarian documentation, and enabling downstream longitudinal analysis of structural damage, similar to efforts seen in remote sensing for disaster relief (Wang et al., 2024).

The dataset provided for this task is uniquely challenging, consisting of images collected from Instagram and published by Palestinian content creators and journalists in Gaza between October 2023 and December 2025. Unlike standardized benchmark datasets, these images are sourced from real-world social media environments characterized by extreme variability in lighting conditions, sensor quality, and camera viewpoints.

The task is formulated as a binary classification problem with two labels: *destruction* and *not_destruction*. Several factors complicate this distinction. First, background clutter, such as rubble-filled streets adjacent to intact buildings can confuse standard classifiers. Second, partial structural damage and occlusions caused by smoke or dust require the model to capture fine-grained visual features rather than relying on coarse global textures. Finally, the inherent class imbalance in archival data necessitates a robust training strategy to avoid bias toward the majority class (Johnson and Khoshgoftaar, 2019).

In this paper, we present a system based on a pretrained ConvNeXt-Tiny backbone, a modernized convolutional architecture that incorporates design principles from vision transformers (Liu et al., 2022). Our methodology emphasizes robustness through four key pillars:

- **Transfer Learning:** Leveraging high-quality feature hierarchies from ImageNet-1k pretraining (Deng et al., 2009).
- **Stratified Ensembling:** Using 5-fold cross-validation to ensure the model generalizes across different data distributions.
- **Class Balancing:** Implementing a weighted cross-entropy loss to penalize the misclassification of the destruction class more heavily.
- **Test-Time Augmentation (TTA):** Enhancing inference-time stability by averaging predictions across multiple image orientations, a technique proven to improve performance in noisy visual domains (Shanmugam et al., 2021).

Our final system achieves a Macro F1-score of 0.8952 and an accuracy of 0.9055 on the official test set. These results demonstrate that modern

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 221–225

convolutional architectures, when combined with careful ensembling and augmentation strategies, provide strong performance for destruction detection in complex social media imagery.

2. Related Work

Image classification has evolved substantially over the past decade, progressing from early deep convolutional neural networks (CNNs) to architectures that integrate design principles inspired by transformers. Residual connections introduced in ResNet enabled the effective training of very deep networks, establishing a strong foundation for large-scale visual recognition. More recently, research has explored incorporating transformer-style components, such as large receptive fields and modern normalization strategies, into convolutional frameworks.

ConvNeXt (Liu et al., 2022) represents a practical extension of this line of work. By modernizing standard convolutional blocks with larger kernels, inverted bottlenecks, and GELU activations, ConvNeXt narrows the performance gap between CNNs and Vision Transformers while preserving the spatial inductive biases inherent to convolutions. Pretraining on large-scale datasets such as ImageNet (Deng et al., 2009) further enhances its suitability for transfer learning on domain-specific tasks with limited annotated data.

2.1. Optimization, Regularization, and Ensembling

Fine-tuning pretrained models on moderate-sized datasets requires careful optimization. AdamW (Loshchilov and Hutter, 2017a), with its decoupled weight decay, has become a standard choice for stabilizing training and improving generalization. Learning rate scheduling strategies such as cosine annealing with warm restarts (Loshchilov and Hutter, 2017b) provide smooth decay and periodic exploration, which can be beneficial in short training cycles typical of shared tasks.

Ensemble learning has long been recognized as an effective technique for reducing model variance and improving generalization (Dietterich, 2000). Aggregating predictions from models trained on different data splits is particularly advantageous when the dataset size is limited or when input data exhibits high variability, as is common in social media imagery.

2.2. Destruction Detection in Crisis Contexts

Automated building damage classification has been widely studied in disaster response and post-conflict assessment, particularly using aerial and satellite imagery (Wang et al., 2024; Al Shafian and Hu, 2024). These approaches typically focus on distinguishing between intact and damaged structures using CNN-based models trained on relatively structured visual inputs.

In contrast, social media imagery presents additional challenges, including extreme viewpoint variation, motion blur, compression artifacts, and inconsistent framing. Research in crisis informatics suggests that robust training strategies and inference-time augmentation can help mitigate noise in user-generated content (Shanmugam et al., 2021). Furthermore, class imbalance is a recurring issue in archival datasets, where intact structures often outnumber damaged ones. Weighted loss functions and related strategies have been shown to improve balanced performance under such conditions (Johnson and Khoshgoftaar, 2019).

Our work builds upon these lines of research by combining a modernized convolutional backbone with stratified cross-validation and probability-level ensembling, tailored specifically to the variability of social media imagery in conflict-related archives.

3. System Description

This section describes our data preprocessing pipeline, model architecture, training strategy, and inference procedure. The architecture of our proposed system follows a robust pipeline consisting of high-resolution preprocessing, a modernized convolutional backbone, and a multi-stage ensemble strategy. A high-level overview of the inference architecture is illustrated in Figure 1.

3.1. Data Preparation and Augmentation

The dataset provided for this task consists of 2,001 high-resolution images sourced from social media. The distribution of classes is inherently imbalanced, reflecting the real-world nature of archival documentation: 706 images (35.3%) are labeled as *destruction*, describing various levels of structural damage, while 1,295 images (64.7%) are labeled as *not_destruction*, showing intact buildings or non-structural scenes.

We combine the provided training and validation splits into a single development pool to maximize the data available for learning while maintaining rigorous evaluation through stratified 5-fold cross-validation. Stratification is critical in this context to preserve the class ratios (*destruction* vs. *not_destruction*) across all folds, ensuring that the model's performance metrics are not skewed by distribution shifts.

All images are resized to 384×384 pixels to preserve fine-grained structural details necessary for

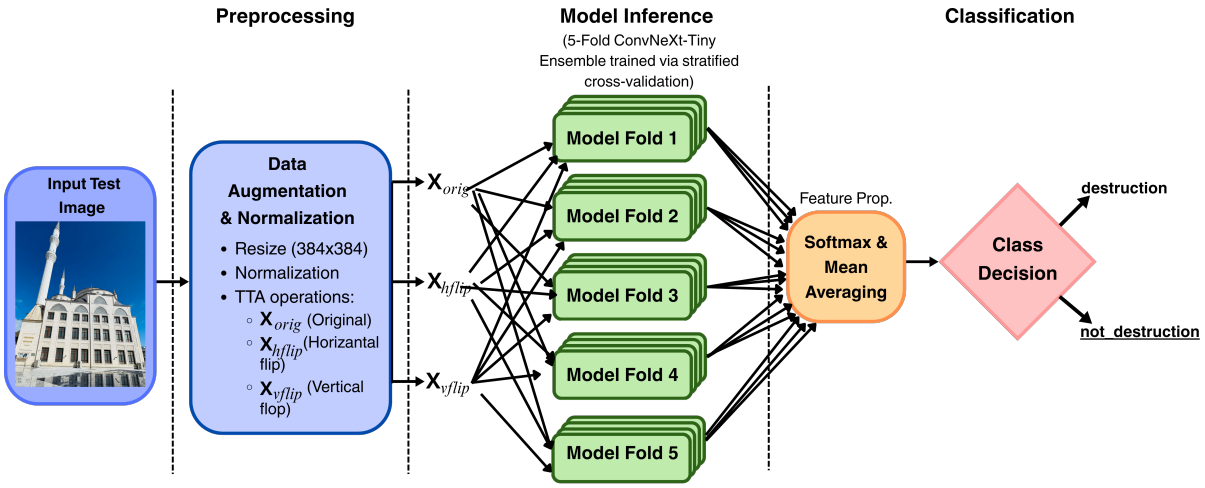

Figure 1: Inference framework of the proposed 5-fold ConvNeXt-Tiny ensemble with test-time augmentation (TTA). Augmented inputs are processed by fold-specific models, and predictions are combined through softmax-based mean averaging to produce the final class label.

damage detection. During the training phase, we apply a stochastic augmentation pipeline to improve generalization, including random horizontal flipping and random rotations. Images are normalized using the ImageNet-1k mean and standard deviation (Deng et al., 2009).

3.2. Backbone Architecture

We employ the ConvNeXt-Tiny architecture (Liu et al., 2022) as our primary feature extractor. ConvNeXt was selected due to its modernized design, which incorporates elements from Vision Transformers (ViTs), such as larger kernel sizes, inverted bottlenecks, and Layer Normalization, while retaining the inductive bias and computational efficiency of fully convolutional networks. We initialize the backbone with weights pretrained on ImageNet-22k and fine-tuned on ImageNet-1k. The final 1000 class linear head is replaced with a custom binary classification head consisting of a single linear layer producing two logits corresponding to the task labels.

3.3. Training Strategy

We train the model using stratified 5-fold cross-validation. For each fold, a separate model is trained and the best checkpoint is selected based on Macro F1-score on the validation split. Furthermore, the model is optimized using the AdamW optimizer (Loshchilov and Hutter, 2017a) with an initial learning rate of 3×10^{-5} and a weight decay coefficient of 0.05 to prevent overfitting on the relatively small dataset.

To mitigate the inherent class imbalance, we utilize a weighted cross-entropy loss function. Based on the class distribution in the training set, we assign a weight of 1.84 to the *destruction* class and 1.0 to *not_destruction*. This weighting scheme forces the model to prioritize sensitivity toward structural damage, which is the minority class in the provided archive.

3.4. Ensemble and Test-Time Augmentation

After training, we obtain five models corresponding to the five folds. During the inference phase, we adopt a dual-level ensemble strategy to maximize stability. First, we apply Test-Time Augmentation (TTA), generating three distinct views for every input image: the original view, a horizontal flip, and a vertical flip. Second, these three views are passed through all five fold-specific models (the best checkpoint from each fold). Let $P_{f,v}$ be the softmax probability vector for fold $f \in \{1 \dots 5\}$ and view $v \in \{1 \dots 3\}$. The final prediction $\hat{y}$ is computed by averaging the softmax probabilities across all 15 permutations (5 folds $\times$ 3 TTA views), with the final label determined by selecting the class with the highest mean probability:

$$\hat{y} = argmax \left(\frac{1}{15} \sum_{f=1}^{5} \sum_{v=1}^{3} P_{f,v} \right) \quad (1)$$

This strategy effectively reduces individual model variance and accounts for orientation differences in archival social media photography.

3.5. Implementation Details

The system is implemented using the PyTorch framework and the `timm` (PyTorch Image Models)

library. All experiments were conducted on a single NVIDIA T4 GPU using Google Colab. The code and pretrained weights will be made available upon acceptance.

4. Results and Discussion

The performance of the proposed ensemble system was evaluated using the official shared task metrics: Macro F1-score and Accuracy. Although Accuracy reflects overall correctness, model selection during cross-validation was based primarily on Macro F1-score, ensuring balanced consideration of both the *destruction* and *not_destruction* classes under potential class imbalance.

Table 1 reports the final performance on the official held-out test set.

Metric	Score
Macro F1-score	0.8952
Accuracy	0.9055

Table 1: Performance of the proposed system on the official test set.

The system achieves strong performance across both metrics, with a Macro F1-score of 0.8952 and an Accuracy of 0.9055. Performance remained consistent across folds, suggesting that the stratified 5-fold setup helped reduce variance due to data partitioning. Averaging predictions at the probability level further improved robustness on unseen data.

For completeness, we also evaluated two alternative model configurations. A single ConvNeXt-Tiny model trained on the original training/validation split achieved a Macro F1-score of 0.8864, highlighting the benefit of the stratified 5-fold ensemble in reducing variance and improving robustness. Another configuration incorporating AutoAugment and label smoothing yielded a Macro F1-score of 0.8904, slightly lower than the proposed approach. These comparisons emphasize that careful class weighting, cross-validation, and test-time augmentation are more effective than simply adding advanced augmentations or smoothing strategies.

Several factors likely contributed to the superior performance of the proposed system. First, the ConvNeXt-Tiny backbone effectively captures multi-scale structural patterns typical of damaged buildings. Its convolutional design preserves spatial inductive biases, which are beneficial for recognizing localized destruction cues such as fractured walls, exposed interiors, and scattered debris in 384×384 images.

Second, Test-Time Augmentation (TTA) improved prediction stability for images collected from mobile devices. Social media imagery often exhibits inconsistent framing and orientation; applying horizontal and vertical flips reduces sensitivity to these variations and encourages more consistent predictions.

Overall, the combination of transfer learning, stratified cross-validation, probability-level ensembling, and inference-time augmentation provides a reliable and reproducible baseline for destruction detection in social media archives, and clearly outperforms simpler or alternative training strategies.

5. Conclusion

This paper presented our submission to the Nakba Image Classification Shared Task at Nakba-NLP 2026. We developed a binary image classification system based on a fine-tuned ConvNeXt-Tiny backbone, trained using stratified 5-fold cross-validation and combined through probability-level ensembling. Test-time augmentation was applied during inference to improve robustness to orientation and framing variations commonly observed in social media imagery.

The final system achieved a Macro F1-score of 0.8952 and an Accuracy of 0.9055 on the official held-out test set. These results indicate that modern convolutional architectures, when paired with careful training and evaluation strategies, remain effective for destruction detection in noisy, real-world settings. In particular, class-weighted optimization and cross-validated ensembling played an important role in maintaining balanced performance across both classes.

Overall, the proposed approach provides a reproducible and competitive baseline for automated destruction classification in archival social media collections.

6. Limitations and Future Work

Despite strong performance, several limitations remain. First, the dataset size is relatively modest, which may limit generalization to rare or atypical visual patterns of structural damage. The model may struggle with subtle cases where damage is partially visible or visually ambiguous. Second, images collected from social media vary substantially in resolution, compression artifacts, and viewpoint, which can introduce unpredictable failure cases.

In addition, the current system is purely visual. It does not leverage accompanying textual captions, hashtags, or temporal metadata, which may provide complementary contextual signals. Incorporating multimodal information could improve robustness, particularly in borderline cases.

Finally, the task formulation is binary and does not capture varying degrees of structural damage. Extending the framework to multi-class or regression-based damage severity estimation

would allow for more fine-grained analysis and may better reflect the complexity of real-world destruction patterns.

Future work will therefore focus on multimodal fusion strategies and on evaluating generalization across datasets collected from different time periods or geographic regions.

7. Ethical Considerations

The dataset used in this study consists of social media imagery documenting structural destruction in Gaza between 2023 and 2025 (Abrahams et al., 2026). The dataset documents structural destruction during an ongoing conflict, and we recognize the sensitivity of working with such material. As the workshop guidelines require, our work focuses on the technical task of automated structural destruction classification to support documentation and archival analysis. We have taken care to ensure that our model operates only at the image classification level and does not extract or infer identifiable personal information. Furthermore, we acknowledge the potential for dual-use of such technologies; while our goal is humanitarian documentation and archival accessibility, automated damage detection systems must be used responsibly to avoid the mischaracterization of civilian infrastructure.

8. Conflict of Interest

The author declares that they have no financial or personal relationships with other people or organizations that could inappropriately influence or bias the results and interpretations presented in this paper.

9. Acknowledgements

The author would like to thank the organizers of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) for providing the dataset and establishing this shared task. We also express our gratitude to the Palestinian journalists and content creators whose documentation work on the ground made this research possible.

10. Bibliographical References

Sultan Al Shafian and Da Hu. 2024. Integrating machine learning and remote sensing in disaster management: A decadal review of post-disaster building damage assessment. *Buildings*, 14(8):2344.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of big data*, 6(1):27.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ilya Loshchilov and Frank Hutter. 2017a. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ilya Loshchilov and Frank Hutter. 2017b. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*.

Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. 2021. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1214–1223.

Lili Wang, Jidong Wu, Youtian Yang, Rumei Tang, and Ru Ya. 2024. Deep learning models for hazard-damaged building detection using remote sensing datasets: A comprehensive review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:15301–15318.

11. Language Resource References

Alexei Abrahams, Shadi Abudalfa, Mustafa Jarrar, and George Mikros. 2026. The nakbaarchive-classifier 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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.