Free-Gaza at NakbaArchiveClassifier Shared Task: Towards Distinguishing the Destructive Effect of Nakba: NakbaImage Classification using Artificial Intelligence Techniques

Nisreen I. R. Yassin, Enas A. Hakim Khalil

Systems & Information Department, National Research Centre (NRC), Giza, Egypt.
nisreen.yassin20@gmail.com, enaskhalil@gmail.com

Abstract

The accounts of the continuing Palestinian Nakba encompass considerable significance. Over the course of the three years of the conflict, millions of photos from social media have been preserved. The preservation and classification of these data through artificial intelligence tools are essential to guarantee their availability, accessibility, and applicability. This paper presents a highly optimized, resource-constrained machine learning pipeline for binary image classification. The system is designed for the NakbaArchiveClassifier Shared Task 2026, which aims to distinguish between images of destroyed infrastructure and those of intact infrastructure. The system uses two lightweight EfficientNetB0 networks to build a weighted ensemble. With strict hardware limitations of 2GB GPU VRAM, the system achieves an F1-score of 84.16%, ranking 9th on the leaderboard.

Keywords: Nakba-NLP 2026, ensemble system, binary image classification

1. Introduction

Following October 7, 2023, a huge number of infrastructure images were shared on social media, documenting the situation in the Giza strip. Visual testimonies provide undeniable documentary evidence of conflict, displacement, and structural impact. The huge volume of unstructured images presents a substantial bottleneck for historians, journalists, and researchers. Manual classification of these images is not scalable, which motivated the development of automated methods for damage detection and classification.

Automating the classification of these visual archives is a crucial step in rendering historical data machine-understandable. By developing robust computer vision models capable of analysing and indexing archival images, researchers gain scalable retrieval tools. This technological intervention not only ensures the long-term accessibility of these narratives for future generations but also allows researchers to efficiently validate and quantify this visual evidence of historical and ongoing destruction within large-scale digital repositories.

To address the challenges of classifying unstructured images, the NakbaArchiveClassifier Shared Task 2026 shared task is formulated as a binary classification problem (Abrahams et al., 2026). Participants must process raw input images and accurately classify each sample image into one of two classes: destruction or non-destruction.

The task introduces many domain-specific challenges:

- Moderate class imbalance (35% positive class)
- High intra-class variance (varied structural types, perspective)
- Social media compression artifacts and heterogeneous capture devices (visually subtle or partial structural damage)

Our Free-Gaza system uses transfer learning EfficientNetB0, fine-tuned on the provided dataset. It is evaluated through accuracy and F1-score. The proposed system provides an efficient and reproducible baseline for classifying images of infrastructure damage collected from social media.

2. Background

Disaster response monitoring has extensively studied infrastructure damage detection. For post-disaster assessment, early methods primarily relied on remote sensing and satellite imagery. Convolutional neural networks (CNNs) showed strong performance in recognizing structural damage patterns, such as collapsed building roof failure. Recently, the research has expanded to include ground-level imagery shared on social media platforms. The data collected from such platforms is real-time but highly unstructured visual data. Unlike satellite images, social media images exhibit substantial domain variability, including diverse viewpoints, varying lighting conditions, compression artifacts, and heterogeneous capture devices. These factors present significant intra-class variance and make binary classification more challenging, especially when damage is partial or visually subtle.

CNNs have become the dominant method for image classification following the success of deep learning architectures such as AlexNet (Krizhevsky et al., 2012) and ResNet (He et al., 2016). These models learn hierarchical feature representations that capture both low-level

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 133–136

textures and high-level semantic structures. Therefore, these models are well-suited for identifying infrastructure damage. In the case of limited data, transfer learning has shown particular effectiveness. By initializing networks with weights pretrained on large-scale datasets such as ImageNet (Deng et al., 2009), models can leverage general visual priors and adapt them to specific-domain tasks through fine-tuning. This method reduces overfitting and improves convergence when labelled data are scarce. The EfficientNet family (Tan and Le, 2019) introduces compound scaling of network depth, width, and resolution to achieve improved trade-offs between accuracy and efficiency. EfficientNetB0, the baseline variant, achieves competitive performance while using significantly fewer parameters than earlier architectures. Such parameter efficiency is advantageous in shared-task environments and moderate-sized datasets, where overfitting is a concern.

While most destruction detection research focuses on contemporary satellite imagery, the methodologies provide a strong foundation for classifying destruction in historical archives. YOLO-based object detection models have been successfully applied to identify destroyed buildings in high-resolution imagery (Zhang et al., 2023). Semantic segmentation models such as hybrid U-Net architectures enable pixel-level classification of damaged structures (Xu et al., 2022). In data-scarce contexts, unsupervised radar-based approaches have demonstrated the feasibility of detecting building destruction without extensive manual labeling (Kaufmann et al., 2024).

3. System Overview

The proposed image classifier was designed to achieve the best performance on complex, unstructured images while overcoming the stringent hardware limitations of 2GB VRAM. The system shifts from an empirical risk-minimization approach to a data-driven ensemble-based approach. ConvNeXt and DenseNet121 are two examples of networks that are deeper than others and achieve better texture recognition. However, empirical evidence during development revealed that these networks' performance was decreased because they had to use smaller batch sizes. To extract as many feature representations as possible from low-resolution 224 x 224-pixel input, the system uses EfficientNetB0 as its global backbone. EfficientNet's compound scaling algorithm will uniformly scale the network's width, depth, and resolution (Tan and Le, 2019).

Transfer learning is used to train EfficientNetB0 using pre-trained weights from ImageNet (Deng et al., 2009). Model A is trained on the original data using standard binary cross-entropy. This caused the model to over-optimize for "Not-destruction" buildings that were easy to identify. At the same time, it struggled with ambiguous landscapes after a disaster, such as partial structural collapse or rubble hidden behind other buildings. Model B is trained using the original training data and a set of external data. The external data includes the AIDER dataset (Kyrkou and Theocharides, 2020; Kyrkou and Theocharides, 2019), pseudo-labels data, which are predictions exhibiting extreme statistical confidence, and the old validation data. For Model B, the learning objective was changed to binary focal cross-entropy (Lin et al., 2017). This algorithm reduces the impact of class imbalance by scaling the loss based on the model's confidence in its predictions and adapting to the varying levels of difficulty across the examples in the dataset. For the enriched dataset, a learning rate is scheduled using a cosine decay learning rate (Loshchilov and Hutter, 2016). The approach entails a smoothly and continuously reduced learning rate over time. The non-linearly decaying toward convergence allowed Model B to converge securely in 12 Epochs.

In the prediction phase, the 8-Way Test-Time Augmentation (TTA) Algorithm (Shorten and Khoshgoftaar, 2019) is implemented. For each test image, the algorithm generates a set of eight augmented views, including the original, a horizontal flip, a vertical flip, rotations of 90, 180, and 270 degrees, and adjustments to brightness and contrast. The model evaluates all eight views and calculates the final prediction probability as the arithmetic mean of the outputs. A weighted soft voting mechanism is used for final classification (Zhou, 2025). Model A is trained purely using original data, while model B is trained using original data and AIDER data; therefore, the ensemble probability is calculated as follows:

$$P_{ensemble} = 0.40P_{model\ A} + 0.60P_{model\ B}$$

4. Experimental Setup

The proposed system depends on two datasets. The first dataset is the original dataset provided by NakbaArchiveClassifier Shared Task 2026. This dataset contains 1599 PNG images, divided into 1400 for training, 199 for validation, and 402 for testing. The dataset demonstrates moderate class imbalance and high visual variability, including disparities in lighting, perspective, resolution, and compression artifacts characteristic of social media platforms. The second dataset is the AIDER (Aerial Image Dataset for Emergency Response Applications). The construction of the AIDER dataset entailed the manual collection of all images about four distinct disaster events. This collection encompassed 521 images of fire and smoke, 526 images of flooding, 511 images of collapsed buildings and rubble, and 485 images of traffic accidents. In addition, 4,863 images were collected for the normal case. The data structure

is summarized in Table 1. Data usage for models A and B is illustrated in Table 2. For replication of the 0.8416 F1-score of NakbaArchiveClassifier Shared Task 2026, the hyperparameter configurations utilized for models A and B are depicted in Table 3. Each of these two models employs an alternative optimization approach to support its distinct functions in the ensemble. While Model A aims to achieve steady, background-level convergence, Model B aggressively selects hard examples from the augmented dataset. The development, training, and inference phases of NakbaArchiveClassifier were built on open-source software. The system relies on standard Python machine learning libraries and the TensorFlow framework.

Classes labels	original training data	Original val. data	AIDER data	Test data
Destruction	494	70	2043	-
Not destruction	906	129	4863	-
total	1400	199	6906	402

Table 1: Dataset structure

	Training	Validation
Model A	Original dataset	Original dataset
Model B	Original (494+906) + AIDER (2043+1600) + Pseudo-labels + Old Val.	Split 10% of the training data.

Table 2: Data usage for model A and model B

Parameter	Model A	Model B
Base architecture	EfficientNetB0	EfficientNetB0
Weights	ImageNet	ImageNet
Resolution	224x224x3	224x224x3
Batch size	32	16
Dropout Rate	0.3	0.4
Classification head	Dense 1 unit sigmoid	Dense 1 unit sigmoid
Voting weight	0.4	0.6

Table 3: Hyperparameter configurations utilized for models A and B

5. Results

Several experiments were performed to evaluate the proposed system. The experiments were conducted on an Intel® Core™ I7 CPU at 2.60 GHz computer using the TensorFlow platform for machine learning. The performance of the proposed system is evaluated on the validation set following each training epoch. The evaluation metrics used include accuracy, macro F1-score, precision, recall, and a confusion matrix.

Model A has been trained on the original data from NakbaArchiveClassifier and evaluated on the validation set of 199 PNG images using standard binary cross-entropy. As shown in Table 4, the validation data contain two classes: destruction (70 images) and not-destruction (129 images), indicating class imbalance. Therefore, the model is biased toward predicting the images as not destroyed. The recall for the destruction class is 0.67, indicating poor classification of destroyed images, with 23 destroyed buildings missing. The confusion matrix of model A is shown in Figure 1.

	Precision	Recall	F1-score	support
destruction	0.80	0.67	0.73	70
Not-destruction	0.84	0.91	0.87	129
accuracy			0.82	199
Macro avg.	0.82	0.79	0.80	199
Weighted avg.	0.82	0.82	0.82	199

Table 4: Classification report of model A

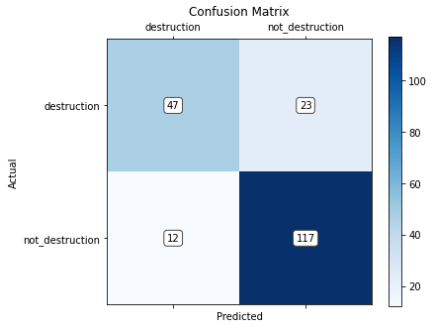

Figure 1: Confusion matrix of model A.

Model B has been trained on a combination of data: original data, AIDER data, pseudo-labels (200 images), and old validation data (199 images). In the validation phase of model B, the validation data has been increased to 1090 images, selected at random from the AIDER data. Table 5 shows the classification report of model B. Its confusion matrix is shown in Figure 2. According to the classification report for model B,

the validation data contains 581 destructive images and 509 non-destructive images. The confusion matrix shows the sensitivity of model B to the destruction images, with a recall of 0.97. This behavior reflects the use of Focal Loss. The overall macro F1-score of model B is 0.85, which is greater than the macro F1-score of model A of 0.80. The ensemble system combines the two models. The system uses a 40/60-weighted soft-voting mechanism and a test-time augmentation mechanism. This system successfully balanced precision and recall across the two classes. The final achieved macro F1-score is 0.8416 on the unlabeled official test set. The ensemble system's classification report is shown in Table 6.

	Precision	Recall	F1-score	support
destruction	0.79	0.97	0.87	581
Not-destruction	0.96	0.71	0.82	509
accuracy			0.85	1090
Macro avg.	0.87	0.84	0.85	1090
Weighted avg.	0.87	0.85	0.85	1090

Table 5: Classification report of model B

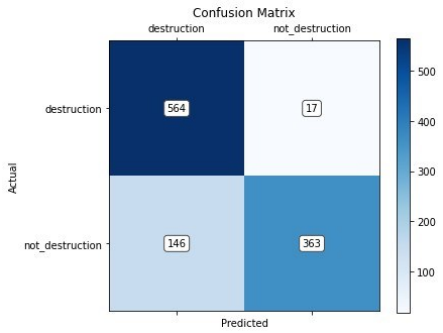

Figure 2: Confusion matrix for model B.

accuracy	0.8557
precision	0.8427
recall	0.8405
F1-score	0.8416
specificity	0.8405
Balanced-accuracy	0.8405

Table 6: Classification report of the ensemble system

6. Conclusion

The development of the NakbaArchiveClassifier pipeline suggests the potential for achieving highly competitive classification metrics. The proposed classification system successfully functioned under significant hardware limitations (a 2-gigabyte virtual memory limit). The system considered both the data and learning objectives to be dynamic variables. Therefore, it was able to overcome severe domain shifts and class imbalance. The evaluation of the proposed ensemble system shows promising results. For future work, our model's capabilities could be further refined with more powerful hardware and additional training data.

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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp. 248-255.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pp. 2980-2988.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983.

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. Journal of big data, 6, pp. 1-48.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, PMLR, pp. 6105-6114.

Zhi-Hua Zhou. 2025. Ensemble methods: foundations and algorithms. Chapman and Hall/CRC.