“Hope” at NakbaArchiveClassifier Shared Task: Transfer Learning-Based CNN Models for Infrastructure Damage Detection

Lojien AlKhidir, Heba Abdelhady

Hamad Bin Khalifa University
Doha, Qatar
loal89338@hbku.edu.qa, heab89248@hbku.edu.qa

Abstract

This paper describes Team Hope’s system for the NakbaArchiveClassifier Shared Task at Nakba-NLP 2026. The task focuses on binary classification of social media images into two categories: destruction and not_destruction. We evaluated multiple convolutional neural network architectures using transfer learning, including ResNet34, ResNet50, EfficientNet-B0, and a fine-tuned ResNet34 variant with staged training. All models were initialized with ImageNet pretrained weights and fine-tuned on the provided dataset of 2,001 images. The dataset is moderately imbalanced and contains visually diverse Instagram images depicting intact and damaged infrastructure. Our best-performing model, ResNet34 trained for 25 epochs with Adam optimizer and a learning rate of 1e-4, achieved 81% accuracy on the evaluation platform. We provide a comparative analysis of the tested architectures and discuss the impact of model depth, training duration, and class imbalance. Given the political and ethical sensitivity of the dataset, we also include a discussion of responsible AI considerations and potential limitations. Our findings suggest that moderate-depth architectures can generalize effectively in low-resource, contextually complex visual classification tasks (Abraham, et al. May 2026).

Keywords: image classification, infrastructure damage detection, transfer learning, convolutional neural networks, Nakba-NLP

1. Introduction

The NakbaArchiveClassifier Shared Task (Abrahams et al., 2026) aims to classify social media images into two categories: destruction and not_destruction. The task contributes to organizing large-scale visual archives related to infrastructure conditions in Gaza.

Automated image classification can support large-scale documentation and indexing of visual material. However, the dataset presents several challenges, including limited size, class imbalance, compression artifacts from social media platforms, and contextual ambiguity in damage representation.

As a multidisciplinary team newly entering computer vision research, we focused on evaluating established convolutional neural network (CNN) architectures using transfer learning (Shin, et al. 2016). Our goal was to compare different pretrained models and analyse their behaviour on a relatively small and context-sensitive dataset.

2. Related Work

Infrastructure damage detection has been widely studied in disaster response contexts such as earthquakes and floods. CNN-based models have demonstrated strong performance in visual damage classification tasks due to their ability to capture hierarchical image features (He, et al. 2016)

Transfer learning from ImageNet-pretrained models is common practice for small and medium-sized datasets, as it allows models to leverage generalized visual representations (Deng et al., 2009; Shin et al., 2016). Prior research suggests that moderate-depth architectures often generalize better than very deep models in low-resource settings.

3. Dataset and Task Description

The dataset provided for the NakbaArchiveClassifier Shared Task consists of 2,001 Instagram images manually labeled into two binary categories: *destruction* and *not_destruction*. The images depict infrastructure scenes collected from social media and include buildings, streets, residential areas, and urban environments under varying visual conditions.

During the development phase, the organizers provided predefined splits:

- **Training set:** 1,400 images
- **Validation set:** 199 images
- **Test set:** 402 images

We strictly adhered to the official splits and did not perform any reshuffling or cross-validation to ensure consistency with the evaluation protocol.

The dataset exhibits a noticeable class imbalance, with approximately 35% of images labeled as *destruction* and 65% labeled as *not_destruction*. This imbalance increases the risk of biased predictions toward the majority class and may influence model behavior during training. In particular, models trained without class weighting may tend to overpredict the *not_destruction* class.

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 187–190

The images vary significantly in resolution, camera angle, lighting conditions, and compression artifacts due to their origin on Instagram. Some images depict clear structural collapse, while others contain partial or subtle damage. In certain cases, both intact and damaged structures appear within the same frame, introducing contextual ambiguity.

The evaluation metric used by the shared task platform was accuracy, calculated as the proportion of correctly classified test images. While accuracy provides a straightforward measure of overall performance, it does not fully reflect per-class performance in imbalanced datasets, where alternative metrics such as macro-F1 could provide additional insight (Sokolova and Lapalme 2009).

4. Methodology

4.1 Preprocessing

All images were resized and normalized using standard ImageNet preprocessing statistics (mean and standard deviation) (Den, et al. 2009). We replaced the final fully connected layer of each pretrained network with a new linear layer containing two output neurons corresponding to the binary labels (*destruction* and *not_destruction*).

We used the official train, validation, and test splits provided by the organizers without modification to ensure comparability with other participating systems. Data loading was implemented using PyTorch's DataLoader, with batch sizes of 32 for ResNet models and 16 for EfficientNet-B0.

No additional manual data cleaning or resampling was performed. For EfficientNet-B0, class imbalance was addressed through weighted loss.

4.2 Model Architectures

We evaluated four convolutional neural network architectures initialized with ImageNet pretrained weights. In all cases, the final classification layer was replaced to adapt the models to binary classification.

4.2.1.1 ResNet34

- Pretrained on ImageNet
- Final fully connected layer replaced (2 output units)
- Batch size: 32
- Optimizer: Adam
- Learning rate: 1e-4
- Epochs: 25
- Loss: CrossEntropyLoss

4.2.1.2 ResNet50

- Pretrained on ImageNet
- Final fully connected layer replaced (2 output units)
- Batch size: 32
- Optimizer: Adam
- Learning rate: 1e-4
- Epochs: 10
- Loss: CrossEntropyLoss

4.2.1.3 EfficientNet-B0

- Pretrained on ImageNet
- Final classification layer replaced (2 output units)
- Batch size: 16
- Optimizer: Adam
- Learning rate: 1e-3
- Epochs: 25
- Loss: Weighted CrossEntropyLoss (to mitigate class imbalance)

4.2.1.4 ResNet34 Advanced

- Pretrained on ImageNet
- Final fully connected layer replaced (2 output units)
- Two-stage training strategy:
    - First 5 epochs: backbone frozen, training only final layer
    - Next 5 epochs: full fine-tuning
- Initial learning rate: 1e-3 (final layer)
- Fine-tuning learning rate: 1e-4
- Scheduler: ReduceLROnPlateau
- Batch size: 32

All models were trained using the Adam optimizer (Adaptive Moment Estimation) and standard backpropagation (Kingma and Ba, 2015). The best-performing model on the validation set was selected for final test submission.

5. Experimental Results

We evaluated four pretrained convolutional neural network architectures under consistent training conditions. Model selection was based on validation performance, and final predictions were submitted to the shared task evaluation platform.

Overall, the results indicate relatively close performance across architectures, with accuracy ranging between 79% and 81%. Despite architectural differences in depth and training strategy, performance variation remained within a narrow margin, suggesting that dataset size and class imbalance may constrain achievable gains.

ResNet34 achieved the highest performance, indicating that moderate-depth architectures may generalize better in small-scale datasets compared to deeper networks such as ResNet50. While ResNet50 contains more parameters and higher representational capacity, it did not provide measurable improvement under the current training configuration.

EfficientNet-B0, trained with weighted cross-entropy loss to address class imbalance, achieved comparable but slightly lower performance than ResNet34. This suggests that class weighting alone may not substantially alter model behavior when the imbalance is moderate (Tan and Le 2019).

The advanced ResNet34 configuration incorporating staged freezing and learning rate scheduling did not improve results. The shorter training duration and higher initial learning rate may have limited its convergence potential.

5.1 Accuracy scores of evaluated architectures

The results demonstrate that stable fine-tuning with moderate learning rates and sufficient training epochs contributed more to performance gains than architectural depth or complexity.

Model	Accuracy
ResNet34	81%
ResNet50	80%
EfficientNet-B0	80%
ResNet34 Advanced	79%

Table 1: Accuracy scores of evaluated architectures

6. Error Analysis

Analysis of misclassified samples revealed several recurring patterns. The most common errors included:

- Partial structural damage misclassified as intact when destruction affected only a limited region of the image.
- Distant or small-scale destruction embedded within otherwise intact urban scenes.
- Visual noise and reduced clarity due to Instagram compression artifacts.
- Ambiguous scenes containing both intact and damaged structures, leading to contextual confusion.

A significant proportion of false negatives occurred when damage did not visually dominate the frame. In such cases, the model appeared to rely on global scene characteristics rather than localized damage cues. This suggests that the architectures may prioritize overall texture and structure patterns rather than fine-grained spatial irregularities.

Additionally, class imbalance may have influenced prediction behavior, as models tended to favor the majority *not_destruction* class in borderline cases. This tendency is consistent with the higher prevalence of intact scenes in the training data.

These findings indicate that binary classification of infrastructure damage is not purely a visual clarity problem but also a contextual interpretation challenge. Future improvements may benefit from incorporating attention mechanisms or region-based approaches that focus on localized structural anomalies rather than entire-image representations.

7. Lessons Learned

As a multidisciplinary team newly entering the field of computer vision, this project provided several practical and methodological insights.

First, we learned that architectural complexity does not necessarily guarantee superior performance. Despite evaluating deeper and more sophisticated models, a moderately sized architecture (ResNet34) trained with stable hyperparameters achieved the best results. This highlighted the importance of controlled experimentation over architectural novelty.

Second, we observed that hyperparameter choices such as learning rate, number of training epochs, and class weighting significantly influence performance, particularly in small and moderately imbalanced datasets. Careful adjustment of these parameters proved more impactful than increasing model depth.

Third, we gained a deeper understanding of the challenges posed by real-world social media imagery. Unlike curated benchmark datasets, Instagram images exhibit compression artifacts, varying resolutions, and contextual ambiguity, which complicate binary classification.

Finally, we recognized the importance of critical reflection in machine learning research. Performance metrics alone do not fully capture system behavior, particularly in sensitive domains. Understanding failure patterns and ethical implications is essential for responsible deployment.

This experience reinforced the value of systematic experimentation, transparent reporting, and cautious interpretation of model results.

8. Conclusion

We presented Team Hope's system for the NakbaArchiveClassifier Shared Task, evaluating four pretrained convolutional neural network architectures under consistent experimental conditions. Among the tested models, ResNet34 achieved the highest performance with 81% accuracy on the evaluation platform.

Our results demonstrate that in small and moderately imbalanced datasets, stable training configurations and appropriate hyperparameter selection can have a greater impact on performance than architectural depth alone. The relatively narrow performance gap across models also suggests that dataset characteristics, including visual variability and contextual ambiguity, impose inherent limitations on achievable accuracy.

Beyond quantitative performance, this project emphasized the importance of systematic experimentation, careful interpretation of results, and ethical awareness when working with sensitive visual data. Automated classification systems in such domains should be viewed as supportive analytical tools rather than definitive decision-making systems.

9. Ethical Considerations and Limitations

The dataset used in this shared task is politically and contextually sensitive. Automated classification systems in such domains raise concerns regarding potential misuse, narrative framing, and bias amplification. The dataset originates from specific social media sources, which may introduce representation bias in terms of geography, perspective, and visual framing.

Additionally, the labeling of destruction versus non-destruction may involve subjective interpretation, particularly in cases of partial or ambiguous damage. Such subjectivity can influence model learning and evaluation outcomes.

The relatively small dataset size and class imbalance further limit generalizability. Models trained on this dataset may not transfer reliably to other geographic regions, conflict contexts, or image sources with different visual characteristics.

There is also a risk of automation bias, where outputs from machine learning systems may be interpreted as objective or authoritative despite their inherent limitations. For this reason, the model should not be interpreted as a definitive tool for real-world humanitarian assessment but rather as an experimental research system intended for controlled academic evaluation.

We emphasize the importance of transparency, responsible reporting, and cautious deployment when working with sensitive visual data.

10. 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 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, 770–778.

Mingxing Tan, and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, 6105–6114.

Marina Sokolova, and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. In Information Processing & Management 45 (4), 427–437.

Hoo-Chang Shin, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, I. Nogues, Jianhua Yao, Daniel Mollura, and Ronald M. Summers. 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. In *IEEE Transactions on Medical Imaging* 35 (5), 1285–1298.