# No Overfit at NakbaArchiveClassifier Shared Task: A Swin Transformer-Based System for Destruction Image Classification

**Mohamed F. Mohamed[1,2], Samar M. Abd El-Mageed[1,2], Ensaf H. Mohamed[1,3]**
[1]Computer Science Department, Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt
[2]Computer Science Department, Faculty of Computers and Information, Luxor University, Luxor, Egypt
[3]School of Information Technology and Computer Science, Nile University, Giza, Egypt
mfathy@fci.luxor.edu.eg, samar.mahmoud@fci.luxor.edu.eg, enmohamed@nu.edu.eg

**Abstract**

Automated destruction identification from visual data plays a critical role in large-scale documentation, humanitarian analysis, and digital archiving of conflict-related events. Within this context, the Nakba-NLP 2026 Workshop introduced a shared task aimed at training and evaluating a binary image classification model to distinguish between destroyed or damaged infrastructure and intact infrastructure. However, the limited dataset size and the visual variability of real-world scenes make this task particularly challenging. This work presents a Swin Transformer–based framework tailored for destruction image classification. The proposed model employs a hierarchical Swin Transformer backbone for robust feature extraction, followed by a multi-layer perceptron classifier for decision-making. To address the limited data issue, transfer learning and a customized training strategy are applied to adapt the model effectively without full end-to-end retraining. Furthermore, a semi-supervised data expansion approach is utilized to enlarge the training set from 1,400 to 10,000 images, improving model generalization and robustness. Experimental results on the official blind test set demonstrate strong performance, achieving an F1-score of 86.55% and an accuracy of 87.81%, ranking 5th in the shared task.

**Keywords:** image classification, deep learning, Swin transformer, multi-layer perceptron network, transfer learning

## 1. Introduction

Visual content is now an important tool for documenting real-world events and preserving historical memory. Images are essential in humanitarian and conflict-related contexts because they help document infrastructure damage, facilitate large-scale analysis, and support documentation efforts. The NakbaArchiveClassifier (Abrahams et al., 2026) is a shared task that aims to automatically distinguish between destruction and non-destruction images under a realistic evaluation setting. The task presents several challenges. First, it is difficult to train deep neural networks from scratch due to the limited size of the available labeled dataset. Second, robust feature extraction is challenging due to the visual variability of scenes, which includes various viewpoints, lighting conditions, degrees of damage severity, and background clutter. Third, the evaluation is conducted on a blind test set through the CodaBench [1] platform, requiring models that generalize well to unseen data.

To address these challenges, this paper proposes a Swin Transformer–based classification framework with a training strategy tailored to the shared task environment. Instead of training from scratch, pretrained ImageNet-1K weights are utilized to leverage rich visual representations learned from large-scale data. To prevent overfitting and preserve pretrained knowledge, the backbone is frozen, and only a task-specific head block is trained. The head incorporates a Swin stage followed by a multi-layer perceptron (MLP) classifier. Furthermore, to mitigate dataset size limitations, a semi-supervised data expansion strategy is employed to increase the training set using pseudo-labeling and manual verification.

The main contributions of this work can be summarized as follows:

- A Swin Transformer–based model with an MLP classifier head and customized training strategy is proposed for destruction image classification.
- A semi-supervised data expansion approach is employed to increase the training dataset.
- Experimental results on the official test set demonstrate that the proposed classification system achieves strong and balanced performance, reaching an F1-score of 86.55% and an accuracy of 87.81%, and ranking 5th in the shared task [2].

[1] https://www.codabench.org/competitions/12654/#/pages-tab

[2] https://www.codabench.org/competitions/12654/#/results-tab



*Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026*, pages 171–176

## 2. Background

Image classification is a fundamental task in computer vision that aims to assign a predefined label to an input image based on its visual content. Over the years, classification techniques have evolved significantly, progressing from traditional handcrafted feature-based approaches to deep learning-based methods that automatically learn feature representations. The emergence of deep learning, particularly convolutional neural networks (CNNs), revolutionized image classification. Architectures such as AlexNet(Krizhevsky et al., 2012), VGG(Simonyan and Zisserman, 2015), ResNet(He et al., 2016), and EfficientNet(Tan and Le, 2019) demonstrated the ability to learn hierarchical spatial features directly from raw pixel data.

Although CNN-based models have a strong capability to capture local patterns, they may struggle to effectively model global context due to their limited receptive fields. To address this limitation, transformer-based architectures have been introduced into vision tasks. Vision Transformers (ViT)(Dosovitskiy et al., 2021) and their variants employ self-attention mechanisms to model long-range dependencies across image patches. Among these variants, the Swin Transformer (Liu et al., 2021) has demonstrated significant improvements over the original ViT by employing a hierarchical structure with shifted window-based self-attention, enabling scalable and computationally efficient modeling of both local and global information.

Consequently, the Swin Transformer is utilized as the core component of the proposed framework. The Swin Transformer study provides multiple architectural variants that differ in size and computational complexity. In this paper, the tiny variant is selected, as it is more suitable for the dataset constraints.

## 3. Task Description and Dataset

The NakbaArchiveClassifier shared task (Abrahams et al., 2026) is part of the Nakba-NLP 2026 Workshop [3], itself part of LREC 2026 [4]. The task is to develop a binary image classification system that can distinguish between images of damaged or destroyed infrastructure and those of intact infrastructure.

The dataset consists of 2,001 Instagram images posted by Palestinian content creators and journalists in Gaza for documenting events since October 7, 2023. These images are labeled into two classes (destruction or non-destruction) and are divided into training, validation, and testing splits. The details of these distributions are summarized in Table 1.

Table 1: Main Nakba dataset configuration

| Split | Label | # Samples | Label Ratio | Split Size | Split Ratio |
|---|---|---|---|---|---|
| Train | destruction | 494 | 35.29 | 1400 | 69.97 |
| | not_destruction | 906 | 64.71 | | |
| Validate | destruction | 70 | 35.18 | 199 | 9.95 |
| | not_destruction | 129 | 64.82 | | |
| Test | destruction | 142 | 35.32 | 402 | 20.09 |
| | not_destruction | 260 | 64.68 | | |
| **Full Dataset Size** | | **2001** | | | |

## 4. System Overview

### 4.1. Enhanced Dataset

Although the main Nakba dataset provides valuable visual documentation, the size of the training set, which contains 1,400 images, is too small to effectively train a deep learning model. The situation becomes even worse when the classifier employs a transformer-based architecture that typically requires large-scale data. To address this challenge, we employed a semi-supervised training strategy to expand the training set, as illustrated below:

1. A Swin Transformer-based classification model with pretrained weights was trained on the main dataset for three epochs and achieved 86.43% accuracy on the validation set.

2. Several source datasets, which focus on general damage and are partially related to our task, were utilized to increase the number of training images. Table 2 provides an overview of these datasets.

3. The trained model was then used to classify the newly added images into destruction and non-destruction.

4. High-confidence predictions ($> 0.9$ or $< 0.1$) were retained as pseudo-labels, while the remaining samples were manually re-annotated to improve data quality.

Using this strategy, a total of 8,600 images were added to the main dataset, bringing the training set to 10,000 images. The final training set configuration is presented in Figure 1.

### 4.2. Model Architecture

The proposed model adopts a Swin Transformer architecture for destruction image classification and

[3] https://sina.birzeit.edu/nakba-nlp/2026/

[4] https://lrec2026.info/

Table 2: Images data sources

| Dataset | Description | Size |
|---|---|---|
| (Tuncer, 2023) | It contains images representing different types of damage, including the following classes: (1) debris, (2) damaged building, (3) damaged highway, (4) non-damaged building, and (5) non-damaged highway. These images were collected from five online sources and are stored in JPEG or JPG format with varying dimensions. | 4,373 images |
| (Niloy et al., 2021) | This dataset focuses on disasters and contains images classified into one of the following categories: (1) damaged infrastructure, (2) fire disaster, (3) human damage, (4) land disaster, (5) water disaster, and (6) non-damage. | 13,600 images |
| (sta, 2025) | The dataset contains images of war-related events, where each image is in JPG format and is classified into one of the following categories: (1) combat, (2) destroyed, (3) fire, (4) humanitarian, and (5) military. | 1,700 images |
| (Alam et al., 2020) | It is a consolidated benchmark dataset for social media image classification in disaster response. Its images were collected from sources (Alam et al., 2018) and (Mozannar et al., 2018), and have the purpose of being generally employed for four main tasks. The first task focuses on disaster types, the second on image informativeness, the third on humanitarian categories and actions, and the fourth on damage severity detection. | 77,350 images |

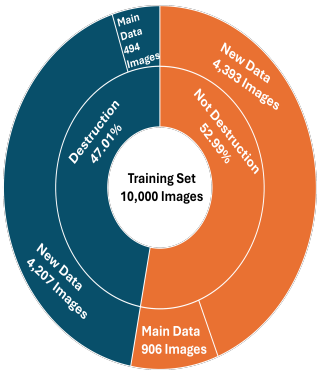


Figure 1: New train dataset configuration

consists of two key components: a backbone block and a head block. Figure 2 presents an overview of the architecture and processing pipeline of the proposed model.

The backbone block is primarily used for feature extraction by employing the first three stages of the Swin Transformer architecture. Specifically, given an input image of size 3×224×224, the backbone first applies patch partitioning followed by linear embedding to convert the image into non-overlapping token representations. The network then processes these tokens through three stages consisting of multiple Swin Transformer blocks (2, 2, and 6 blocks, respectively), with patch merging operations between stages to progressively reduce spatial resolution while increasing feature dimensionality. Each Swin Transformer block consists of two main components: a Window Multi-Head Self-Attention (W-MSA) module and a Shifted Window Multi-Head Self-Attention (SW-MSA) module, both wrapped with layer normalization and residual connections, and followed by a feed-forward MLP module. These window-based attention mechanisms capture local patterns that progressively integrate into a broader contextual representation as the channel depth increases from 96 to 384.

The second component is the head block, whose objective is to classify the input image using the extracted features. Before the final classification process, the head block further refines the representations by employing the last stage of the Swin Transformer architecture, increasing the feature dimension to 768. This strategy customizes the network and allows it to focus more effectively on the target task without requiring full end-to-end retraining, as discussed in Section 4.3. After the final features are condensed via global average pooling and flattening, they are passed to the MLP classification head, which acts as the decision-making module. It consists of two hidden layers with 256 neurons each, followed by a single-neuron output layer. The GELU (Gaussian Error Linear Unit) activation function is employed because it provides smoother nonlinear transformations than ReLU and has demonstrated superior performance in transformer-based architectures. Moreover, dropout layers are incorporated to mitigate overfitting. Finally, the model is trained using the Binary Cross-Entropy (BCE) loss function to optimize the prediction of destruction versus non-destruction labels.

### 4.3. Training Strategy

The training strategy for this model centers on a transfer learning approach, utilizing Swin Transformer weights pretrained on ImageNet-1K to provide a robust foundation for feature extraction. To prevent disrupting these pretrained features and to avoid overfitting due to the relatively limited dataset size, the backbone is frozen and kept non-trainable throughout the training process. In contrast, only the head block is trained to allow the network to specialize in the target task. The head training is performed in two phases: first, the MLP classification module is trained for 30 epochs to stabilize the decision layers and adapt them to the extracted

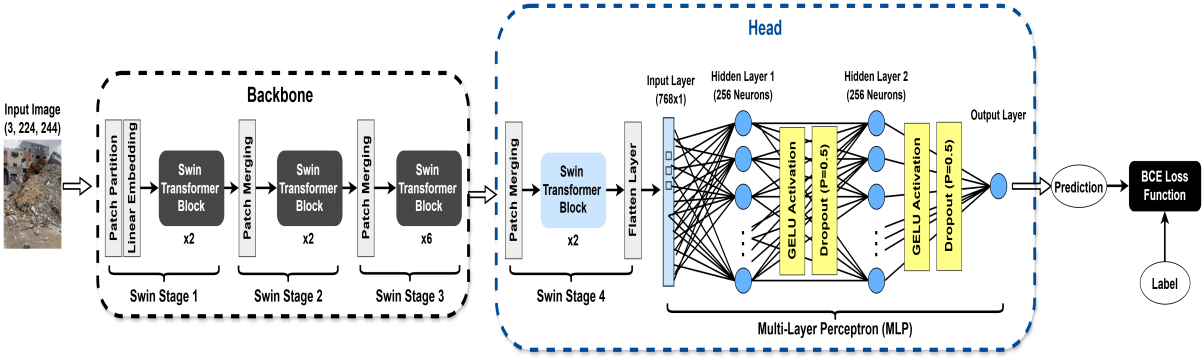


Figure 2: Model architecture

features. Then, the entire head block is trained for an additional 70 epochs to further refine task-specific representations and improve classification performance.

## 5. Experimental Results

### 5.1. Implementation Details

The model was implemented in Python using the PyTorch framework. Training and testing were conducted on the Kaggle platform using a Linux-based environment with a 64-bit x86-64 processor architecture. The system was equipped with a single NVIDIA Tesla P100-PCIE-16GB GPU to enable accelerated computation. The input images have dimensions of 224×224×3 (height × width × channels) and are normalized using the mean and standard deviation of ImageNet(Deng et al., 2009). The AdamW(Loshchilov and Hutter, 2019) optimizer is utilized with default parameters except weight_decay set to 1e-4. A batch size of 32 is adopted for training. The initial learning rate is set to 1e-4 for the first train and 1e-5 for the second train. Primary metrics were used to evaluate and compare the model's performance: (1) Accuracy, (2) Precision, (3) Recall, and (4) Macro F1-score.

### 5.2. Quantitative Results

The proposed model was evaluated on the test set of the NakbaArchiveClassifier Shared Task (Abrahams et al., 2026) by submitting its predictions to the CodaBench platform [5]. It is worth noting that the evaluation was conducted on a blind test set, ensuring strong generalizability validation and fair comparison. Table 3 presents the performance results obtained from the official CodaBench evaluation using multiple metrics.

The model achieved strong classification performance, obtaining an F1-score of 86.55% and an accuracy of 87.81%. The close values between accuracy and F1-score indicate balanced performance across both classes without significant bias. Furthermore, the proposed model ranked 5th among all participating teams in the shared task [6], demonstrating its competitiveness and robustness in comparison with other submitted methods.

Table 3: Model performance on the test set of the Nakba image classification shared task

| Evaluation Metric | Score |
|---|---|
| F1-score | **86.55** |
| Accuracy | **87.81** |
| Precision | **86.88** |
| Recall | **86.26** |

### 5.3. Ablation Analysis

To determine the optimal classification model for destruction identification, suitable for the shared task environment (in terms of dataset characteristics and evaluation metrics), multiple experiments were conducted using different architectural design choices. Table 4 presents the performance results of these experiments on the test dataset using F1-score and accuracy as evaluation metrics.

The reported experiments are those most relevant to the final selected model, where both the backbone and head components were varied in each configuration. Specifically, the backbone options included a transformer-based architecture (Swin), a CNN-based architecture (EfficientNet), and a hybrid CNN–Transformer approach that employs a convolutional block as a preprocessing step followed by Swin as the feature extractor. For the head component, the alternatives included using no additional head (default classifier), an MLP classi-

[5] https://www.codabench.org/competitions/12654/#/pages-tab

[6] https://www.codabench.org/competitions/12654/#/results-tab

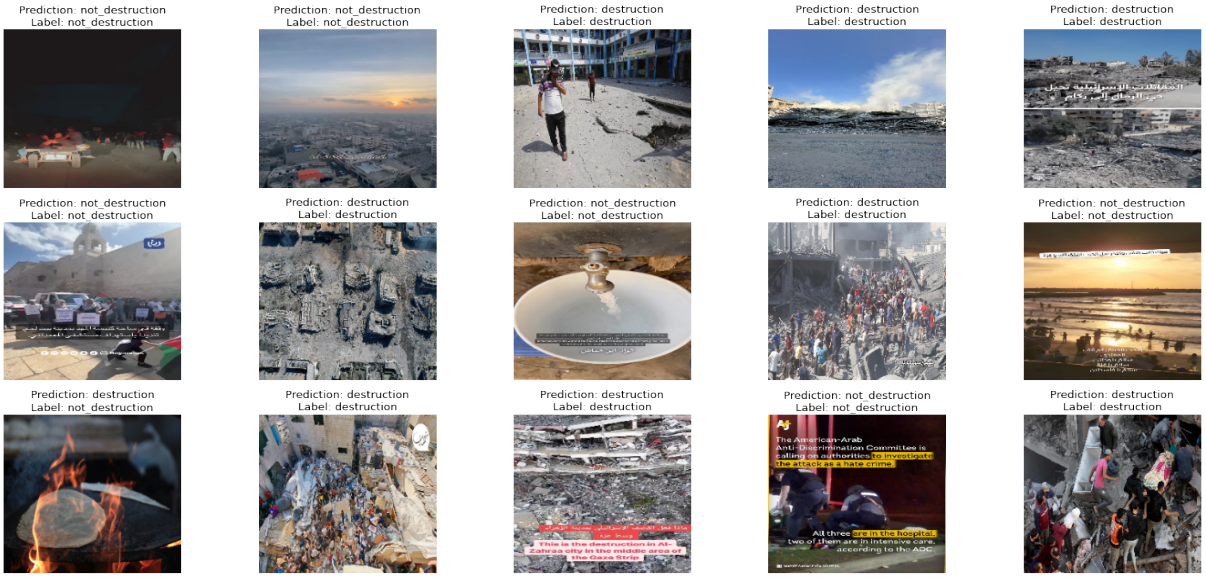


Figure 3: Qualitative results of the proposed framework on the validation set.

Table 4: Ablation analysis of different backbone and head configurations for the Nakba image classification shared task

| Backbone | Head | F1-score | Accuracy |
|---|---|---|---|
| Swin | – | 84.51 | 85.82 |
| Swin | MLP | 86.55 | 87.81 |
| Swin | SVM | 82.03 | 83.58 |
| CNN & Swin | MLP | 83.56 | 84.58 |
| EfficientNet | – | 84.56 | 85.82 |

fier, or a support vector machine (SVM). Overall, the ablation results demonstrate that the Swin Transformer backbone coupled with the proposed MLP head provides the most effective and well-balanced architecture for the target task.

### 5.4. Error Analysis

To better understand the strengths and limitations of the proposed model, an error analysis was conducted by examining the confusion matrix on the validation set, as illustrated in Figure 4.

The model correctly classified 117 non-destruction images and 54 destruction images, demonstrating strong performance across both categories. However, 12 non-destruction samples were incorrectly predicted as destruction (false positives), while 16 destruction samples were misclassified as non-destruction (false negatives). This number of errors indicates that some destruction cases remain challenging, possibly due to subtle damage patterns or visual similarities between damaged and undamaged structures.

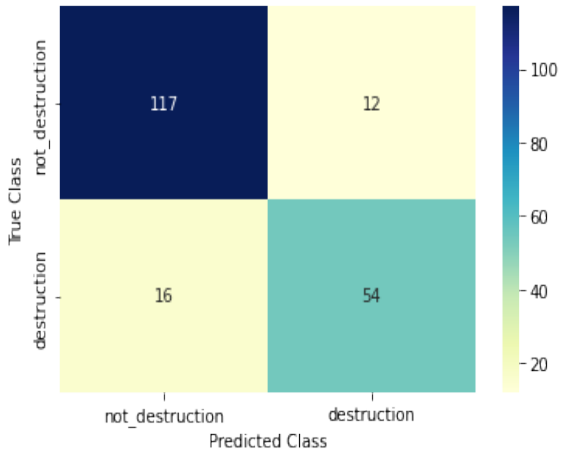


Figure 4: Validation set confusion matrix

Nevertheless, the overall distribution shows that the model maintains a balanced performance between the two classes, without significant bias toward either class.

To qualitatively illustrate the effectiveness of the proposed model, Figure 3 presents representative examples. All these images are randomly sampled from the validation set, which was not used in the training process. Above each image, the true label and the model's prediction are displayed, which highlights the model's ability to capture destruction with a small number of misclassified examples.

## 6. Conclusion

This paper presented a Swin Transformer-based framework for destruction image classification in the Nakba image classification shared task. The

proposed framework achieves strong performance by integrating a Swin Transformer backbone, an MLP classification head, a customized training strategy with transfer learning, and a semi-supervised data expansion strategy. Moreover, extensive ablation analysis demonstrated the effectiveness of the proposed architectural design, confirming that the Swin backbone coupled with the MLP head provides superior and balanced performance. Despite these achievements, the proposed model still struggles with several issues. These issues include that it has not completely addressed overfitting, classification accuracy requires further improvement, and the model's generalizability is still limited. Future work may explore partial backbone fine-tuning, employing advanced data augmentation techniques, further improvements in the dataset (size and diversity), and adopting multi-head classification.

## 7. References


2025. stable dataset. https://universe.roboflow.com/nothing-cniqf/stable-d0jqc. Roboflow Dataset.

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.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*.

Firoj Alam, Ferda Ofli, Muhammad Imran, Tanvirul Alam, and Umair Qazi. 2020. Deep learning benchmarks and datasets for social media image classification for disaster response. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 151–158.

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*, pages 248–255. Ieee.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*. Published as a conference paper at ICLR 2021. Originally submitted to arXiv in October 2020.

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 (CVPR)*, pages 770–778.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, volume 25, pages 1097–1105.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Hussein Mozannar, Yara Rizk, and Mariette Awad. 2018. Damage identification in social media posts using multimodal deep learning.

Fahim Faisal Niloy, Abu Bakar Siddik Nayem, Anis Sarker, Ovi Paul, M Ashraful Amin, Amin Ahsan Ali, Moinul Islam Zaber, AKM Mahbubur Rahman, et al. 2021. A novel disaster image data-set and characteristics analysis using attention model. In *2020 25th International Conference on Pattern Recognition (ICPR)*, page 6116–6122. IEEE.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*. Published as a conference paper at ICLR 2015. Originally submitted in 2014.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, pages 6105–6114.

Türker Tuncer. 2023. Damaged constructions image dataset. https://www.kaggle.com/datasets/turkertuncer/damaged-constructions-image-dataset. Kaggle dataset.