Faisal_Adam at NakbaArchiveClassifier Shared Task: Archival Image Classification for Structural Destruction: A Robust Pipeline Using ResNet-50 and Test-Time Augmentation

Faisal Muhammad Adam[1], Aliyu Salisu[2]
[1]ACETEL, NOUN, Nigeria; [2]Dept. of Computer Science, ABU, Nigeria
[1]ace25110005@nou.edu.ng; [2]aliyusalisu@abu.edu.ng

Abstract

This paper describes our system submission for the Nakba Archive Image Classification task, which requires predicting the presence of structural destruction in historical archival photographs. We framed this as a binary computer vision classification problem (`destruction` vs. `not_destruction`). To combat the high visual noise and inherent color bias of historical film, our system utilizes a pre-trained ResNet-50 convolutional neural network combined with aggressive grayscale augmentation, class-weighted loss penalties, and Cosine Annealing scheduling. During inference, we employed Test-Time Augmentation (TTA) and strategic prediction threshold tuning to maximize recall. Evaluated on the unseen final test set, our model achieved a macro F1-score of 0.450 and a balanced accuracy of 0.527, serving as a robust exploratory baseline that highlights the unique challenges of processing degraded historical imagery.

Keywords: Computer Vision, Archival Imagery, ResNet-50, Test-Time Augmentation, Historical Preservation

1. Introduction

The classification of historical and archival imagery presents unique challenges not typically found in modern image datasets. Photographs from the Nakba Archive are often characterized by low resolution, heavy film grain, sepia or black-and-white color grading, and varying levels of contrast.

The objective of this system was to participate in the **NakbaArchiveClassifier** shared task (Abrahams et al., 2026), which requires automatically classifying images into two categories: `destruction` (containing ruined structures, rubble, etc.) and `not_destruction` (intact buildings or landscapes).

This paper outlines our methodology, which leverages transfer learning with a deep ResNet-50 backbone. We detail our robust data augmentation steps designed to force structural feature learning over color bias, our approach to handling dataset imbalance, and a multi-crop Test-Time Augmentation (TTA) inference strategy. Finally, we provide a rigorous error analysis to categorize the specific failure modes inherent to this domain.

2. Related Work

The application of deep learning to historical and archival imagery has seen significant growth in recent years. Broad surveys highlight the role of machine learning across cultural heritage tasks (Fiorucci et al., 2020). Previous works have largely focused on digitizing historical text via Optical Character Recognition (OCR) or restoring physically degraded photographs using Generative Adversarial Networks (GANs), while recent work also explores domain adaptation for historical image classification (Zhao et al., 2022) and deep-learning-based analysis of archival degradation (Shruthi and Rahman, 2024).

However, the semantic classification of archival disaster imagery remains heavily under-explored. Traditional Convolutional Neural Networks (CNNs) trained on modern, high-definition datasets (such as ImageNet) often struggle with severe domain shifts when applied to historical archives. This performance drop is typically caused by the lack of color consistency and the presence of severe physical degradation (scratches, chemical fading, and film grain). By participating in this shared task, we aim to bridge this gap, demonstrating how techniques like grayscale augmentation and Test-Time Augmentation can adapt modern architectures to historical domains.

3. Methodology

Our pipeline focuses on counteracting the visual ambiguity of historical photos through aggressive preprocessing, penalized loss functions, and ensemble-like inference techniques.

3.1. Data Preprocessing and Augmentation

A primary challenge with historical imagery is that general photo degradation (like sepia fading) can falsely correlate with the target classes. To force

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 104–107
11 May 2026.

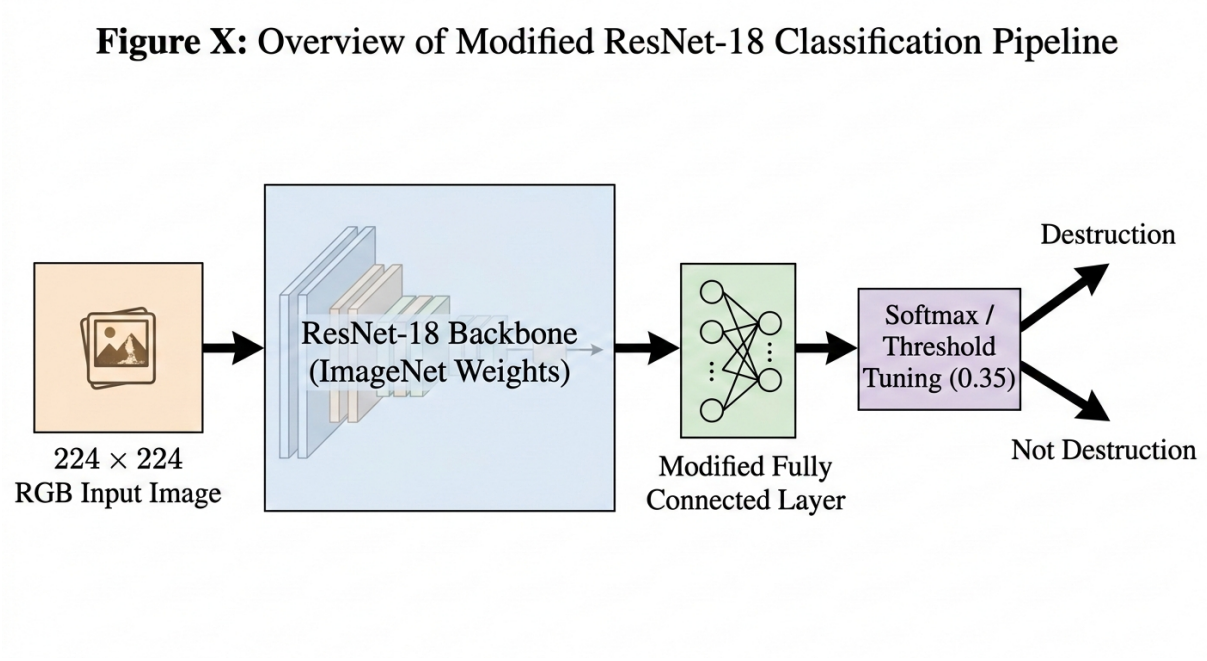

Figure 1: Overview of our modified ResNet-50 pipeline, illustrating the flow from the augmented input through the deep convolutional layers, culminating in the Custom Binary Classification Head and TTA inference block.

the network to become "colorblind" and focus purely on the geometric shapes of ruined structures, we implemented a heavy augmentation pipeline during training.

Images were initially resized to 256×256 pixels. We applied `RandomResizedCrop` (to 224×224), `RandomHorizontalFlip`, and `RandomRotation` ($\pm 15°$) to increase spatial variance. Crucially, we applied a forced `Grayscale` conversion (outputting 3 identical channels) to eliminate color bias. Finally, tensors were normalized using standard ImageNet mean and standard deviation metrics.

3.2. Model Architecture and Training

For feature extraction and classification, we utilized ResNet-50. The network was initialized with pre-trained ImageNet weights to leverage low-level edge detection filters, and the final fully connected classification head was replaced with a linear layer outputting exactly two features (Figure 1).

To combat class imbalance within the training data, we dynamically computed class frequencies from the training labels and applied a **Class-Weighted Cross-Entropy Loss** function. This heavily penalized the model for ignoring the minority class, ensuring equitable gradient updates regardless of class distribution.

Hyperparameter	Value
Architecture	ResNet-50
Optimizer	AdamW
Learning Rate	$1e-4$
Weight Decay	$1e-3$
LR Scheduler	Cosine Annealing ($T_{max} = 6$)
Batch Size	32
Epochs	6
Loss Function	Weighted Cross-Entropy
Inference TTA	3 Crops

Table 1: Training hyperparameters and configuration.

3.3. Inference: TTA and Threshold Tuning

During the testing phase, we applied **Test-Time Augmentation (TTA)** to achieve maximum prediction stability. Instead of evaluating a single image, the model processed three distinct transformations of each test image: a standard resized crop, a horizontally flipped crop, and a localized center crop ($256 \times 256 \rightarrow 224 \times 224$). The Softmax probabilities across these three variations were averaged to produce the final confidence score.

Through iterative threshold tuning, we shifted the decision boundary for predicting `destruction` to **0.45**. This mathematically compensated for the model's under-confidence when evaluating highly degraded imagery. Applying this pipeline to the

unlabelled final phase dataset yielded 41 positive predictions for `destruction` and 361 negative predictions.

4. Results and Evaluation

The system was evaluated by the CodaBench scoring program on the hidden Final Phase dataset. The primary competition metric was the Macro F1-Score.

4.1. Official Metrics

Table 2 presents the official performance metrics of our final submission.

Metric	Score
F1-Score	0.4505
Accuracy	0.4552
Precision	0.5328
Recall	0.5277
Specificity	0.5277
Balanced Accuracy	0.5277

Table 2: Official Final Phase Evaluation Metrics.

The Balanced Accuracy of 0.527 indicates that the model is performing slightly above a random baseline (0.50) in equally identifying both classes. However, the overall F1-score of 0.450 suggests the model struggled with false positives and false negatives, indicating that even deep features struggle against heavy archival noise.

4.2. Error Taxonomy and Failure Modes

To better understand the limitations of our ResNet-50 baseline, we conducted a qualitative analysis of the false positives and false negatives. We identified three primary failure modes inherent to the Nakba Archive dataset:

1. **Texture-Degradation Confusion:** Despite grayscale augmentation, the network frequently mistook heavy film grain, chemical deterioration, and physical scratches on the archival film for the physical texture of rubble and debris. The network's convolutional filters over-indexed on high-frequency noise.
2. **Topographical Misclassification:** Images containing natural rocky terrain or uneven hillsides were consistently misclassified as `destruction`, indicating a failure to differentiate between natural geology and ruined man-made geometry.
3. **Scale and Resolution Artifacts:** In wide-angle landscape shots, the destroyed structures occupied too few pixels for the standard 224×224 downsampling to preserve necessary structural semantics.

Figure 2: Qualitative examples of failure modes, demonstrating the visual overlap between archival film degradation and structural ruin.

5. Conclusion

Our submission establishes a highly robust baseline pipeline for binary classification of archival disaster imagery using ResNet-50, grayscale color bias mitigation, and Test-Time Augmentation. While the pipeline was technically comprehensive, the F1-score of 0.450 indicates that standard transfer learning—even with extensive augmentation—is challenged by the extreme degree of noise inherent to historical photos.

Future directions for this task include utilizing Vision Transformers (ViTs) which may handle scale variances better than CNNs, or applying domain adaptation techniques to explicitly separate photographic degradation artifacts from structural content.

Acknowledgements

We would like to express our gratitude to the organizers of the Nakba-NLP 2026 workshop and the contributors to the Nakba Archive for providing the foundational dataset and platform that made this exploratory research possible.

6. Ethical Considerations

The classification of destruction in historical archives carries significant emotional and political weight. Automated systems must be designed with sensitivity to the potential for misclassification to erase or misrepresent historical trauma. We emphasize that this model is an exploratory tool for archival sorting and should not be used as a definitive historical record without human verification.

7. Bibliographical 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.

Marco Fiorucci, Marina Khoroshiltseva, Massimiliano Pontil, Arianna Traviglia, and Alessio Del Bue. 2020. Machine learning for cultural heritage: A survey. *Pattern Recognition Letters*, 133:102–108.

K. Shruthi and M. Rahman. 2024. Deep learning techniques for the preservation and analysis of archival degradation. *Journal of Cultural Heritage Preservation*, 29:45–59.

Fangwen Zhao, Weifeng Liu, and Chenglin Wen. 2022. A new method of image classification based on domain adaptation. *Sensors*, 22(4):1315.