Brittle-Ductile Fracture Classification Using Deep Learning

Ben Levy

Introduction

Fractography, the analysis of fracture surfaces to determine failure mechanisms, remains a critical but labor-intensive task in materials science. Traditional approaches require trained metallurgists to manually inspect scanning electron microscopy (SEM) images, a process that can take 30 minutes to 2 hours per sample and is inherently subjective. This project explored deep learning approaches for automating the binary classification of fracture surfaces as brittle or ductile, with the goal of improving objectivity, standardization, and throughput in failure analysis.

This was a two-person project completed with Patrick Akers.

Methods

Dataset and Preprocessing

We used a publicly available dataset of 1,818 labeled SEM images of metal fracture surfaces from Zenodo (Campari, 2025), comprising 822 brittle and 996 ductile samples at 224×224 resolution. The dataset was split with 15% held out as a final test set (273 images), with the remaining 1,545 images used for 5-fold cross-validation. All images were normalized to the 0–1 range. For the Vision Transformer experiments, we implemented data augmentation including random flips, rotations, zoom, contrast adjustments, and Gaussian noise to artificially expand the effective dataset size.

Model 1: Custom CNN

Our first approach was a custom CNN built from scratch to establish a baseline. The architecture consisted of four convolutional blocks with batch normalization and max pooling, followed by fully connected layers. The model totaled approximately 19 million parameters, with the vast majority (18 million) concentrated in the flatten-to-dense connection. This was intentionally a simple architecture to get hands-on experience with the data before attempting more sophisticated approaches.

Model 2: ResNet (Transfer Learning)

We implemented a transfer learning approach using ResNet50 pretrained on ImageNet. The pretrained convolutional base was used as a feature extractor with a custom classification head added for our binary task. This approach leverages the rich visual features already learned from millions of natural images, requiring only fine-tuning for our specific domain.

Model 3: Vision Transformer

We were also interested in exploring Vision Transformers (ViT) given their recent success in computer vision tasks. The architecture splits each 224×224 image into 196 patches of 16×16 pixels, projects each patch to an embedding space, adds learnable positional encodings, and processes the sequence through transformer blocks with multi-head self-attention. Our initial implementation (v1) used 128-dimensional embeddings, 4 attention heads, 4 transformer layers, dropout of 0.1, and AdamW optimization with weight decay of 1e-5.

When v1 showed signs of overfitting, we developed v2 with aggressive regularization: reduced projection dimension (64), fewer transformer layers (3), increased dropout (0.3), stronger weight decay (10×), L2 regularization on dense layers, label smoothing (0.1), and a cosine learning rate schedule with warmup. We applied all these changes simultaneously rather than incrementally, a decision that proved instructive.

Model 4: Multimodal LLM Comparison

As a thought experiment, we tested whether general-purpose multimodal LLMs could perform this classification task zero-shot. We provided 10 randomly selected test images to Claude Sonnet 4.5 and Claude Opus 4.5 with no context about fracture patterns,simply asking for brittle/ductile classification. This explored whether these models had internalized relevant visual patterns from their training data.

Results

Custom CNN: The model severely overfit, achieving near-perfect training accuracy (~95%) while validation accuracy plateaued below 50%. The loss curves showed classic overfitting behavior with training loss approaching zero while validation loss diverged. In retrospect, the 18 million parameters in the dense layers were far too many for our dataset size, allowing the model to memorize training examples rather than learn generalizable features.

ResNet: The transfer learning approach achieved approximately 98% test accuracy with strong generalization. The confusion matrix showed 117 true positives for brittle (6 false negatives) and 150 true positives for ductile (0 false negatives). Filter visualization revealed that specific convolutional filters consistently activated for brittle versus ductile patterns, demonstrating that the network learned meaningful visual features. A pruned version using only the top 10 most discriminative filters maintained 97.8% accuracy, suggesting the classification relies on a small set of robust features.

Vision Transformer v1: Cross-validation accuracy was 70.49% ± 8.13%, with test accuracy of 73.26%. While better than random, the model showed clear overfitting - training accuracy reached ~95% while validation plateaued around 70%. The attention map visualizations were interesting but ultimately showed the model was biased toward predicting brittle, suggesting it learned spurious patterns rather than the true distinguishing features.

Vision Transformer v2: The heavily regularized version performed dramatically worse: CV accuracy of 55.15% ± 0.63% and test accuracy of 54.95%—essentially random chance with remarkably low variance. The classification report revealed the model was predicting ductile almost exclusively (93% recall for ductile, only 9% for brittle). By applying all regularization techniques simultaneously, we over-constrained the model to the point where it could not learn anything meaningful.

LLM Comparison: On the 10-sample test, Claude Sonnet 4.5 achieved 60% accuracy, misclassifying 4 brittle samples as ductile. Claude Opus 4.5 achieved 100% accuracy on the same samples. While the sample size is too small for statistical significance, the result suggests that larger multimodal models may have internalized relevant visual patterns for materials analysis. This warrants further investigation with proper validation.

Conclusions and Future Work

This project reinforced several important lessons about deep learning in practice. First, architecture choice matters more than hyperparameter tuning when dataset size is limited. The ResNet's skip connections and pretrained features enabled generalization that our custom architectures could not achieve from scratch on ~1,800 images. Second, Vision Transformers are genuinely data-hungry, lacking the inductive biases of CNNs (local spatial structure, translation equivariance), they must learn these patterns from data, requiring datasets an order of magnitude larger than ours.

The ViT v1 → v2 progression taught us that regularization is not monotonically beneficial. When v1 overfit, our instinct was to apply every regularization technique available. The catastrophic performance drop in v2 demonstrated that we over-corrected—the model's capacity was reduced below what the task required. An incremental approach, testing one technique at a time, would have been more informative.

For future work, a pretrained ViT with transfer learning (analogous to the ResNet approach) would likely perform well while maintaining the interpretable attention mechanisms. The LLM results, while preliminary, suggest an intriguing direction: exploring whether vision-language models could provide not just classification but explanations referencing domain knowledge about fracture mechanisms. Finally, extending to multi-class classification (including mixed-mode fractures) and testing generalization across different materials and imaging conditions would be valuable for practical deployment.

References

Campari, A. (2025). SEM Image dataset (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.15510590

Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021.

He, K., et al. (2016). Deep residual learning for image recognition. IEEE CVPR, 770-778.