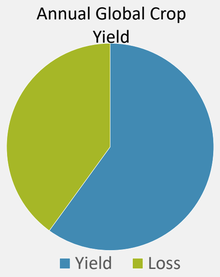
A Transformer-CNN Hybrid Autoencoder for Semi-Supervised Plant Disease Detection

Mulham Fawakherji
College of Science and Technology, North Carolina A&T University

Introduction

Plant diseases pose a major threat to crop productivity and food security. Traditional visual inspection is slow, labor-intensive, and often inaccurate.

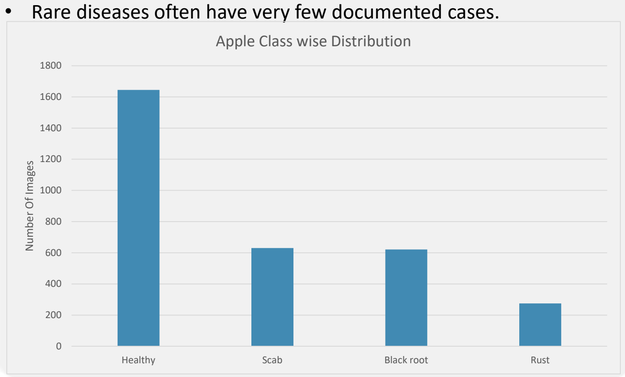
Machine learning offers a fast, scalable, and objective solution by automatically identifying disease symptoms from leaf images. Modern deep learning models such as CNNs and Vision Transformers can learn



subtle visual patterns, enabling early detection and timely intervention. This work presents an automated plant disease detection pipeline designed to improve accuracy, reduce manual workload, and support precision agriculture practices.

Dataset and Challenges

- Healthy crop images are highly abundant.
- Diseased samples are limited, particularly for early-stage infections.
- Certain diseases appear only in specific seasons or geographical regions.
- Rare diseases often have very few documented cases.



Methodology

Dual Feature Extraction:

- A CNN Encoder encoder (ResNet-based) captures local texture and edge features related to disease patterns.
- A Vision Transformer (ViT) encoder divides the image into patches and models global contextual relationships via self-attention.

Feature Fusion:

- CNN and ViT feature maps are spatially aligned and fused (concatenation) to combine local and global information.

Decoder & Reconstruction:

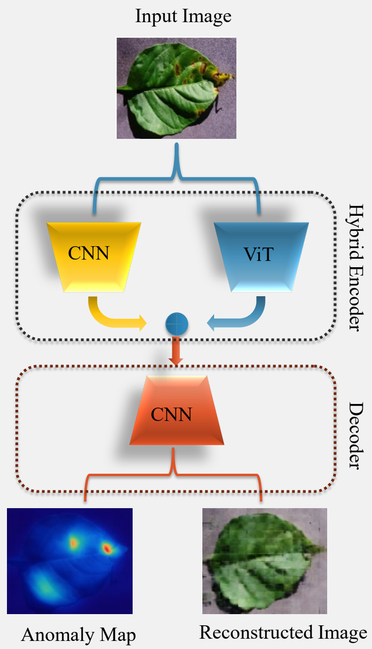
- A transposed-convolution decoder reconstructs the input image from the fused features.

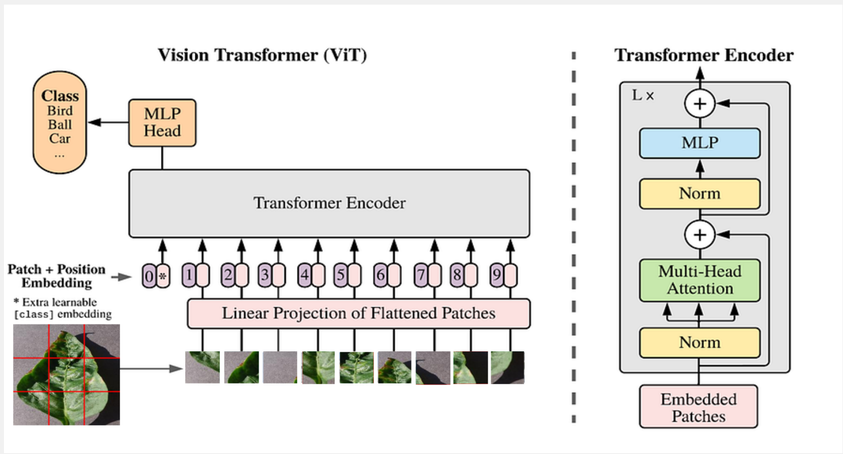
Training Strategy (Healthy-Only Learning):

- The model is trained only on healthy plant images.

Loss Function:

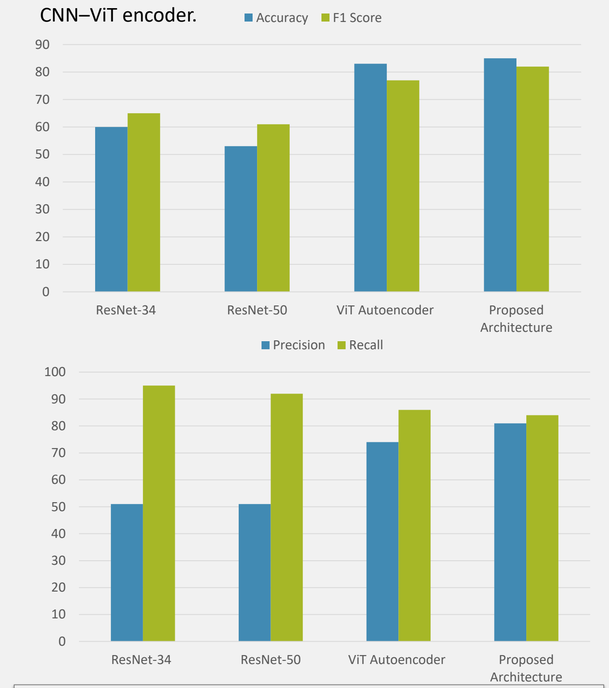
- Training is guided by a weighted combination of Structural Similarity (SSIM) loss and L1 reconstruction loss.





Results

- Performance comparison of different autoencoder architectures using Accuracy, F1 Score, Precision, and Recall. All evaluated models are autoencoder-based, employing different encoder backbones, including CNN-based encoders (ResNet-34 and ResNet-50), a Vision Transformer (ViT) encoder, and a hybrid CNN–ViT encoder.



Conclusions

- ViT-CNN hybrid autoencoder achieves state-of-the-art unsupervised plant disease localization.
- Trains only on healthy leaf images no annotations needed.
- Superior precision-recall balance.
- Scalable, robust, and generalizable across species.