CNN with Feature Attention and Mutual Information for Plant Disease Diagnosis with Treatment Recommendations

Abstract

Plant diseases remain a major threat to global agriculture, contributing to estimated annual crop losses of 20–40% worldwide. Timely and accurate diagnosis is critical to mitigating these impacts and ensuring food security. In this work, we present a novel deep learning framework for plant disease detection that combines a convolutional neural network (CNN) with feature attention and mutual information (MI) regularization. The proposed method enhances feature selection by assigning dynamic importance to discriminative visual patterns, while MI estimation encourages the model to learn more informative and label-relevant representations.

Evaluated on the PlantVillage dataset encompassing 38 disease classes, our model outperforms a standard ResNet-18 baseline, achieving a better improvement in accuracy and greater robustness across challenging disease categories. To ensure interpretability, we incorporate Grad-CAM visualizations that highlight image regions contributing most to the model's decisions, fostering trust and transparency in practical settings. Additionally, we deploy our trained model via an interactive web application that allows users to upload plant leaf images, receive predictions with confidence scores, view attention-based heatmaps, and obtain curated treatment recommendations. This integration bridges the gap between cutting-edge AI research and real-world agricultural usability, empowering farmers and stakeholders with accessible, actionable diagnostic tools.

Introduction:

This project presents a deep learning framework that integrates attention mechanisms and mutual information (MI) regularization within a Convolutional Neural Network (CNN) to enhance the detection of plant diseases. Utilizing the PlantVillage dataset, the model not only improves classification accuracy but also offers interpretability through Grad-CAM visualizations. A web-based interface further facilitates user interaction, allowing for image uploads, disease predictions, and treatment recommendations.

Motivation

Early detection is vital to prevent disease spread and minimize chemical use. An automated system could empower farmers with real-time insights, improving sustainability and productivity. Integrating feature attention and mutual information enhances model precision and reliability.

Why is it important?

Farmer Challenges: Early disease symptoms are hard to distinguish visually, especially in rural or low-resource settings.

Lack of Experts: Most regions lack access to trained plant pathologists for accurate diagnosis.

Why It's Interesting

This work combines classification accuracy with interpretability using Grad-CAM, bridging AI and agriculture. It tackles visual similarity challenges in disease images through attention-guided feature learning. Additionally, we deploy a practical web interface offering diagnosis and treatment recommendations.

Literature Review

Belghazi, M.I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., (2018).
What they did: Proposed a differentiable neural estimator of Mutual Information using the Donsker–Varadhan representation of KL divergence. They presented a Mutual Information Neural Estimator (MINE) that is linearly scalable in dimensionality as well as in sample size, trainable through back-prop, and strongly consistent. And presents a handful of applications on which MINE can be used to minimize or maximize mutual information. Appiled MINE to improve adversarially trained generative models.
What's missing: It was not applied to plant disease or feature attention tasks directly — it lacks domain-specific integration and interpretability on its own.A nd alone MINE is computationally expensive and sensitive to training instability, requiring careful design choices .

Mohanty, S.P., Hughes, D.P., & Salathé, M.. (2016)
What they did: Applied deep learning (CNNs like AlexNet and GoogLeNet) on the PlantVillage dataset, achieving high classification accuracy (~99%) under controlled conditions.
What's missing: Their model lacks interpretability and is limited in generalizability to field conditions due to its reliance on raw accuracy alone. No attention mechanisms or feature selection techniques were explored, and the approach does not scale well without annotated real-world data.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017).
What they did: Introduced Gradient-weighted Class Activation Mapping (Grad-CAM) to visually explain CNN decisions by highlighting important image regions. They combined Grad-CAM with existing fine-grained visualizations to create a high-resolution class-discriminative visualization, Guided Grad-CAM, and apply it to image classification, mage captioning, and visual question answering (VQA) models, including ResNet-based architectures.
What's missing: While Grad-CAM improves interpretability, it doesn't modify the training process or enhance feature selection — it's only post-hoc explanation.

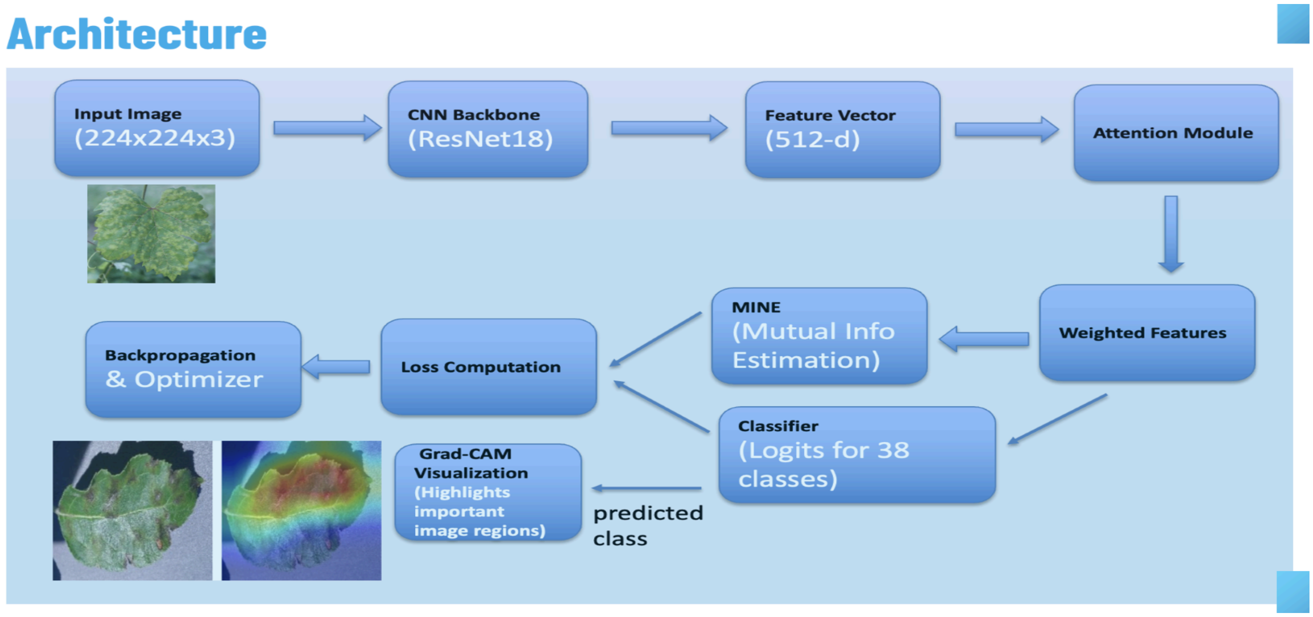
Baseline:

As a starting point, we adopt a **standard ResNet-18 Convolutional Neural Network (CNN)** architecture for image classification using the PlantVillage dataset. This baseline model performs disease prediction without any feature attention or mutual information guidance.

About Data Set:

Contains **over 54,000 images** of healthy and diseased plant leaves.
Covers **38 classes** across **14 crop species**, including tomato, potato, grape, apple, corn, and more.
Images are available in three formats: **color**, **grayscale**, and **segmented** —This project uses the **color images** for richer feature representation. The dataset is partitioned into training, validation, and test sets to evaluate model performance effectively
All images were captured in **controlled conditions** with consistent backgrounds.

Model architecture



Architecture

Input Image (224x224x3)

CNN Backbone (ResNet18)

Feature Vector (512-d)

Attention Module

Weighted Features

MINE (Mutual Info Estimation)

Classifier (Logits for 38 classes)

Loss Computation

Backpropagation & Optimizer

predicted class

Grad-CAM Visualization (Highlights important image regions)

Input Image (224×224×3)

Each input is a colored plant leaf image resized to 224×224 pixels with 3 RGB channels.These images are passed to a convolutional neural network for feature extraction.

CNN Backbone (ResNet18)

ResNet18, pretrained on ImageNet, is used to extract hierarchical features from the image.
The final fully connected (classification) layer is removed, and only the **512-dimensional feature vector** is retained.

Feature Vector (512-d)

This vector contains 512 abstract features learned by the CNN that describe the leaf image.
But not all features are equally important — hence the need for attention.

Attention Module

A learnable neural network that assigns importance weights (0 to 1 via sigmoid) to each of the 512 features.
This step highlights which features the model should "focus" on.

Weighted Features

The original 512 features are **multiplied element-wise by their corresponding attention weights**, producing refined (focused) feature vectors.
This helps suppress irrelevant features and highlight the most discriminative ones.

Mutual Information Estimation (MINE)

The MINE module estimates **mutual information** between the weighted features and class labels.
It encourages the network to retain features that have a strong relationship with the output class — improving generalization and robustness.

Classifier (Logits for 38 Classes)

The weighted features are fed into a linear layer to produce **logits** for all 38 plant disease classes. These logits are used to calculate classification probabilities using softmax.

Loss Computation

The final loss combines **Cross-Entropy Loss** (for classification accuracy) , **Mutual Information Loss** (to encourage informative features) , **Attention Regularization** (to prevent overfitting to a few features) and **Diversity Loss** (to encourage broader feature usage)

Backpropagation & Optimizer

The total loss is used to update parameters in the CNN, attention module, and MINE network through backpropagation.
Optimized using AdamW with learning rate scheduling.

Grad-CAM Visualization (for Interpretability)

Grad-CAM highlights the **regions of the input image** that influenced the prediction.
This makes the model **interpretable and trustworthy** by showing visual cues that led to the decision.

Output: The final output is the predicted disease class, along with Grad-CAM heatmaps for interpretability and optional treatment recommendations via the web interface(Attaching the web demo vedio).

Training Objective

Our training objective combines classification accuracy with mutual information maximization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} - \lambda \cdot \text{MI}(\text{weighted\_features}, \text{labels}) + \alpha \cdot \mathcal{L}_{\text{attention\_reg}} + \beta \cdot \mathcal{L}_{\text{diversity}}$$

Where:

L_CE is the cross-entropy loss for classification
MI(weighted_features, labels) is the mutual information estimate
L_attention_reg is a regularization term on attention weights
L_diversity encourages diversity in feature utilization
λ, α, and β are hyperparameters balancing these objectives

Cross-Entropy Loss

This is the standard classification loss that penalizes incorrect predictions.
It compares the model's output (logits) with the true class labels and Ensures classification performance.
Drives the model to maximize the probability of the correct class.

Mutual Information Estimate

Estimated using the **MINE (Mutual Information Neural Estimation)** network.
Measures how much **information the weighted features contain about the labels**.

Higher MI implies that the model's features are more aligned with the actual class distribution.
*Encourages the model to learn **informative and relevant features**.*
This term is **subtracted** (with negative sign) because in code it's defined as -MI loss, but conceptually, **we are maximizing MI**.
λ → Controls the **strength of MI guidance**

Attention Regularization

A penalty term to **prevent overconfidence** in the attention weights.
Implemented using:

$$\mathcal{L}_{\text{attention\_reg}} = \text{mean}(\log(\text{attention\_weights} + \epsilon))$$

Without this, the attention module might assign extremely high weights to a few features, making it prone to overfitting.(*Regularizes the distribution of attention weights.)*
α → Controls **how strongly to regularize the attention weights**

Diversity Loss

Encourages the model to **utilize a broader set of features** rather than always focusing on the same few.
Implemented using the **negative standard deviation** of attention weights across the batch:

$$\mathcal{L}_{\textbf{diversity}} = -\textbf{mean}(\textbf{std}(\textbf{attention\_weights}))$$

A high standard deviation = varied attention, which is desirable.(*Promotes diverse featureusage across images.)*
β → Controls **how much diversity in feature usage is encouraged**

User-Facing Entry Point

The web interface serves as the **front-end layer** where users can upload plant leaf images directly for diagnosis.

Accessible via browser on desktop or mobile.

Model Inference Backend

Uploaded images are passed to the trained **CNN + Attention + MI model** in real time for prediction.Predictions include: **disease class**, **confidence score**, and **Grad-CAM heatmap**

Remedy Recommendation Module

Based on the predicted disease class, the system fetches **predefined natural and pesticide-based remedies**.
Displayed alongside the diagnosis for immediate action by the user.

PDF Report Generation

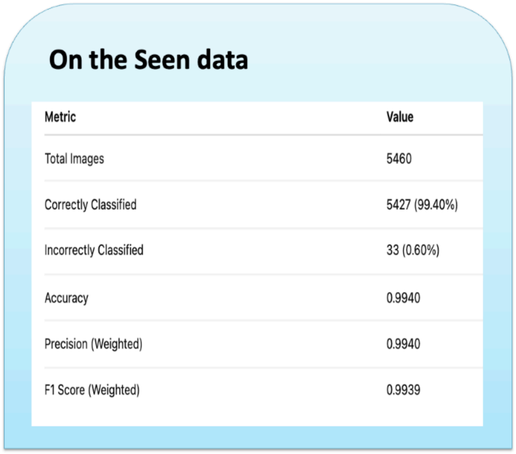
The interface allows downloading a **detailed diagnosis report** (with original image, Grad-CAM, and remedies) in PDF format — useful for field experts or agricultural extension services.

Lightweight and Modular Deployment

The entire web module is built using **Flask**, keeping it light, modular, and easy to deploy on cloud or local devices.

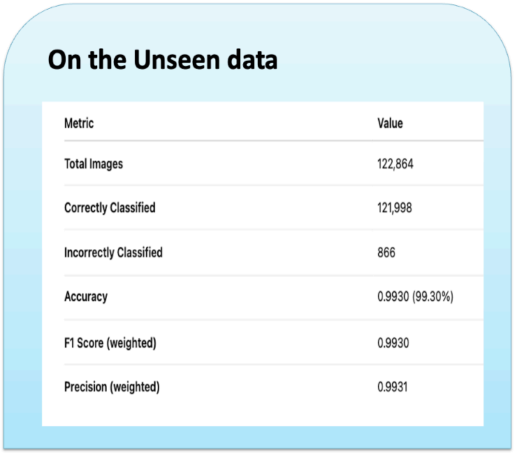
Experimental Results and Analysis

This model was evaluated on both **seen** (test) data and a large set of **unseen** external data. The model generalizes extremely well, maintaining high performance on unseen real-world leaf images, proving robustness



On the Seen data

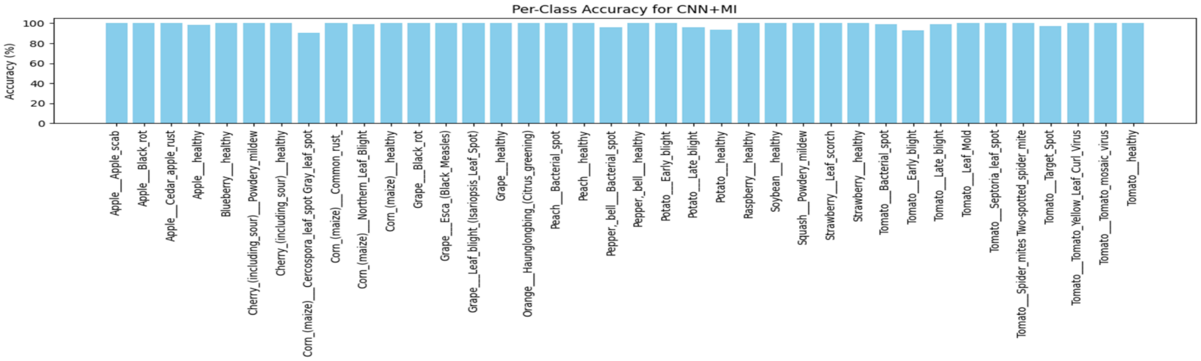
Metric	Value
Total Images	5460
Correctly Classified	5427 (99.40%)
Incorrectly Classified	33 (0.60%)
Accuracy	0.9940
Precision (Weighted)	0.9940
F1 Score (Weighted)	0.9939



On the Unseen data

Metric	Value
Total Images	122,864
Correctly Classified	121,998
Incorrectly Classified	866
Accuracy	0.9930 (99.30%)
F1 Score (weighted)	0.9930
Precision (weighted)	0.9931

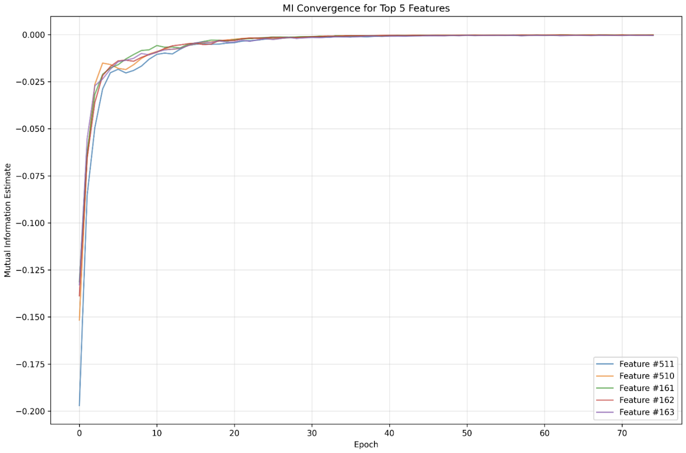
Per-Class Accuracy Analysis: **T**he model performs consistently across all categories, reducing bias toward dominant classes.



- Most disease classes achieve above **95% accuracy**.
- Even visually similar diseases (e.g., Tomato bacterial vs. early blight) show high separability.
- Very few classes fall below 90%, indicating that **feature attention and MI help disambiguate subtle visual cues**.

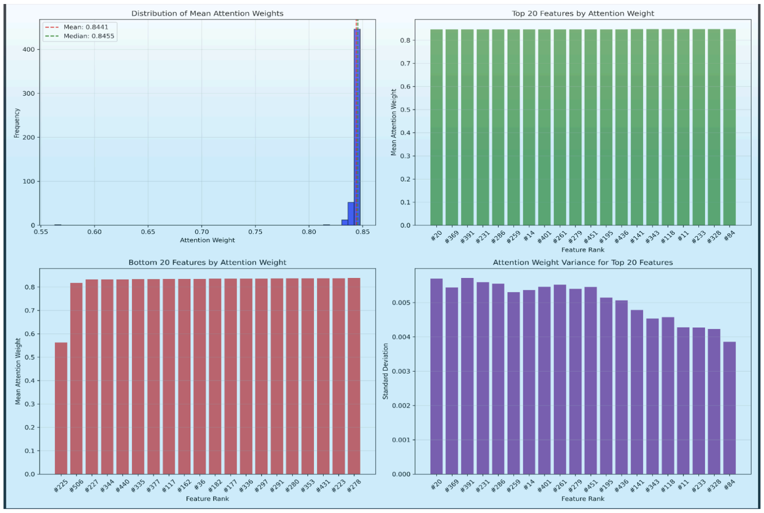
MI Convergence:

This plot shows the convergence of Mutual Information (MI) estimates over training epochs for the top 5 most informative features selected by the model. Initially, all features have low (negative) MI values, but they quickly increase and stabilize around zero, indicating successful optimization. This reflects that the model is learning to align high-MI features with class-relevant information, supporting better discrimination during classification.



Attention-Based Feature Importance Analysis:

This section visualizes how your model **learns which features to focus on** using the attention mechanism. It highlighted key discriminative features, allowing the model to prioritize useful information.



Top-Left: Distribution of Attention Weights
Most weights are concentrated around ~0.84
Suggests strong agreement across features, with few being heavily downweighted.
Top-Right: Top 20 Features by Mean Attention Weight
These features are consistently assigned the highest weights across batches.
These dimensions contribute most to disease discrimination.
Bottom-Left: Bottom 20 Features
Low-weight features contribute the least.
Shows the model's ability to suppress irrelevant or noisy features.
Bottom-Right: Variance in Attention Weights (Top Features)
Shows the stability of attention over training epochs.
Low variance = reliable, confident attention; High variance = inconsistent reliance.

Insight: Feature attention not only improves performance but adds interpretability and regularization, preventing overfitting.

Grad-CAM Visualization:

Grad-CAM generates a **heatmap** that highlights the **regions of the image most responsible** for a particular class prediction.
It involves 3 main steps:

1. Compute Gradients

$$\frac{\partial y^c}{\partial A^k}$$

Compute the gradient of the class score yc with respect to the feature map Ak (from the last convolutional layer).This tells us how much **each feature map channel influences** the class prediction.

2. Global Average Pooling

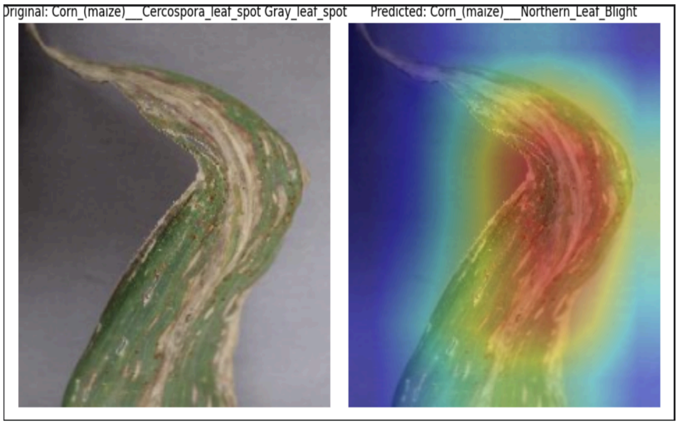
$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

Take the **average gradient value** over all spatial locations (i,j)of each channel.
This gives the **importance weight αkc** for each feature map channel k.

3. Weighted Combination & ReLU

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

Multiply each feature map by its weight and sum across all channels. Apply ReLU to keep Only positive influence
The result is a coarse **heatmap** that highlights important image regions.


Original: Corn_(maize)___Cercospora_leaf_spot Gray_leaf_spot
Predicted: Corn_(maize)___Northern_Leaf_Blight

Left image: The raw input image of a corn leaf showing **Cercospora Leaf Spot**.
Right image: Grad-CAM heatmap overlay showing where the model focused when making its prediction .
Red/yellow areas: Most influential regions.
Blue areas: Least influential regions.

Conclusion and Future Directions

Conclusion:

- Integrating attention mechanisms and mutual information regularization into CNNs significantly enhances the accuracy and interpretability of plant disease detection models.
- To address overfitting, we incorporated **dropout layers** and **attention regularization**, which helped the model generalize better and achieve high accuracy.
- Grad-CAM visualizations further build trust in model predictions by elucidating the decision-making process.
- The development of a web-based interface ensures accessibility, allowing users to leverage the model's capabilities for real-world agricultural applications.

Limitations:

- **Dataset Bias:**T he PlantVillage dataset consists of images captured in controlled environments with uniform backgrounds and lighting.This limits generalization to real-world field conditions, which include complex backgrounds, varying lighting, occlusions, and different imaging devices.
- **Static Treatment Recommendations:** Remedies are fetched from a static remedies.json file, lacking dynamic updates or contextual recommendations.

Future Directions:

- **Domain Adaptation**: Extend the model to handle real-world field images with varying conditions.This will improve robustness against lighting, background clutter, and natural variability in plant appearances.Techniques like adversarial training or domain-invariant feature learning can be explored for adaptation without retraining on large field datasets.
- **Mobile Deployment**: Optimize the model for deployment on mobile devices for on-field diagnostics. Enables real-time, offline disease detection for farmers in remote or low-resource areas.Model compression techniques such as quantization or pruning can help reduce the model’s size while preserving accuracy.
- **Treatment Recommendations**: Incorporate a module suggesting dynamic remedies based on detected diseases.The system can provide location- or severity-specific suggestions using evolving agricultural databasesIntegration with APIs or regional agricultural extensions could enable context-aware, up-to-date guidance.
- **Dataset Expansion**: Include more diverse plant species and disease classes to improve generalization.A larger and more heterogeneous dataset will help the model adapt to global crop varieties and rare diseases.

References

- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.
 In Proceedings of the IEEE ICCV, pp. 618–626.https://doi.org/10.1109/ICCV.2017.74

- Belghazi, M.I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., & Hjelm, R.D (2018).Mutual Information Neural Estimation.In Proceedings of the International Conference on Machine Learning (ICML).https://arxiv.org/abs/1801.04062

- Mohanty, S.P., Hughes, D.P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection.Frontiers in Plant Science, 7, 1419.https://doi.org/10.3389/fpls.2016.01419

- Wang, Y., Zhang, P., & Tian, S. (2024).
 Tomato leaf disease detection based on attention mechanism and multi-scale feature fusion.Frontiers in Plant Science, 15, 1382802.https://doi.org/10.3389/fpls.2024.1382802

- Ma, X., Chen, W., & Xu, Y. (2024).
 ERCP-Net: A channel extension residual structure and adaptive channel attention mechanism for plant leaf disease classification network.Scientific Reports, 14, 4221.
 https://doi.org/10.1038/s41598-024-54287-3

- Poole, B., Ozair, S., van den Oord, A., Alemi, A. A., & Tucker, G. (2019).
 On Variational Bounds of Mutual Information.
 Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97:5171–5180. https://proceedings.mlr.press/v97/poole19a.html

- Gowri, G., Lun, X.-K., Klein, A. M., & Yin, P. (2023).
 Approximating mutual information of high-dimensional variables using learned representations.Proceedings of the National Academy of Sciences (PNAS), 120(8), e2213197120.https://doi.org/10.1073/pnas.2213197120