International Journal of Engineering Applied Science and Management
ISSN (Online): 2582-6948
Vol. 6 Issue 4, April 2025

Fusion of CNN and Transformer-Based Architectures for Enhanced Rice Leaf Disease Detection

Mr. Suganchand Patel[1], Dr. Divyarth Rai[2]

Research Scholar, Department of Computer Science & Engineering, LNCT University, Bhopal (M.P.)[1]
(Associate Professor/ supervisor)) Department of Computer Science & Engineering, LNCT University, Bhopal (M.P.)[2]

Abstract: *Early and accurate detection of rice leaf diseases is vital for sustainable crop management. Conventional CNN-based models, while effective at capturing local features, often fall short in modeling global context, which is crucial in complex field conditions. This paper proposes a fusion architecture that integrates CNNs and Transformer encoders to leverage both local and global feature representations. Through comprehensive experimentation on noisy and clean datasets, the hybrid model demonstrates superior classification performance, robustness, and interpretability compared to standalone architectures.*

Keywords: *CNN, Transformer, Fusion, Classification, Accuracy, Robustness, Attention, Agriculture, Deep Learning, Detection.*

1. Introduction

The agricultural sector faces substantial yield losses due to late or inaccurate diagnosis of plant diseases. Deep learning has transformed plant pathology by offering automatic disease detection through image analysis. Convolutional Neural Networks (CNNs) have dominated this domain due to their capacity to extract spatial hierarchies. However, Transformers—originally designed for natural language processing—have shown promise in vision tasks through self-attention mechanisms. By fusing CNNs with Transformers, we propose a hybrid model that addresses the limitations of CNNs in capturing long-range dependencies and enhances the robustness of rice leaf disease detection. Title: Fusion of CNN and Transformer-Based Architectures for Enhanced Rice Leaf Disease Detection

Agriculture remains the backbone of many developing economies, and rice is one of the most significant staple crops globally. However, rice crops are vulnerable to various leaf diseases that threaten both yield and quality. Traditional methods of disease detection often rely on visual inspection by experts, which is not only time-consuming and labor-intensive but also prone to human error. As a result, the agricultural sector faces substantial yield losses due to late or inaccurate diagnosis of plant diseases. Timely and accurate identification of rice leaf diseases is therefore essential for efficient crop management and minimizing economic loss.

Recent advancements in Artificial Intelligence (AI), particularly Deep Learning (DL), have revolutionized numerous domains, including agriculture. Among DL models, Convolutional Neural Networks (CNNs) have dominated plant disease identification through image analysis. CNNs have demonstrated superior performance in extracting spatial features from leaf images and classifying them into different disease categories. However, CNNs are inherently limited in their ability to capture long-range dependencies and global context information in images. This limitation affects the robustness and generalization capability of CNN-based models, especially when dealing with complex or subtle disease patterns spread across various regions of the leaf.

CNN in Plant Disease Detection

Convolutional Neural Networks have been widely adopted for automated image-based plant disease detection due to their powerful feature extraction capabilities. The architecture of CNNs consists of convolutional layers, pooling layers, and fully connected layers, which hierarchically learn features from images. CNNs such as AlexNet, VGGNet, ResNet, and Inception have shown commendable performance in plant disease classification tasks. For instance, ResNet50 is capable of learning deep representations and has been successfully applied to detect diseases like rice blast, bacterial blight, and brown spot with high accuracy.

However, CNNs process image features in a localized manner. They use fixed-size kernels that slide across the image, which means they excel at recognizing local patterns like edges, spots, and textures. This localized processing leads to limitations when global context is important, such as when disease symptoms are spread over large, non-contiguous regions of the leaf. Additionally, CNNs are less effective at modeling long-distance spatial relationships, which are often critical for accurate classification.

Transformers and Vision Applications

Transformers were originally introduced in the field of Natural Language Processing (NLP) through the revolutionary paper "Attention is All You Need" by Vaswani et al. (2017). The core innovation of the Transformer architecture is the self-attention mechanism, which allows the model to weigh the importance of different parts of the input data and capture global dependencies effectively. This concept was later adapted for computer vision tasks in the form of Vision Transformers (ViTs).

Vision Transformers divide an input image into a sequence of fixed-size patches, which are linearly embedded and processed using Transformer encoders. The self-attention mechanism in ViTs enables them to model relationships across all patches, thereby capturing global contextual information that CNNs may miss. This makes Transformers particularly suitable for complex visual recognition tasks, including plant disease detection, where long-range dependencies between different regions of a leaf are important.

Despite their strengths, Vision Transformers have their own limitations. They require large datasets to train effectively and lack the inductive biases (like locality and translation invariance) that make CNNs efficient on smaller datasets. Therefore, while ViTs excel at capturing global features, they may not perform as well as CNNs in scenarios with limited training data or when fine-grained local features are crucial.

Fusion of CNN and Transformer Architectures

Given the complementary strengths and weaknesses of CNNs and Transformers, recent research has explored hybrid architectures that combine the two. A fusion of CNN and Transformer components can leverage the local feature extraction capabilities of CNNs and the global context modeling ability of Transformers. This approach is particularly promising for rice leaf disease detection, where both local and global patterns are vital for accurate classification.

In a typical CNN-Transformer fusion model, the CNN layers are used to extract low-level and mid-level spatial features from the input image. These features are then passed to a Transformer module, which applies self-attention mechanisms to learn high-level global dependencies among different regions of the image. The fusion may occur at various levels of the network architecture, such as early fusion (after initial CNN layers), mid-level fusion, or late fusion (before the classification layer).

Proposed Hybrid Model for Rice Leaf Disease Detection

The proposed hybrid model begins with a ResNet50 backbone that performs initial convolutional operations on the input image. This step captures local features such as textures, edges, and color patterns indicative of specific rice leaf diseases. The feature maps from the CNN are then flattened and converted into a sequence of tokens suitable for Transformer processing. These tokens are enriched with positional encodings to retain spatial information.

The Transformer block processes the token sequence using self-attention and feedforward layers. This allows the model to understand the relationships between spatially distant regions of the image, capturing global disease patterns and subtle contextual cues. Finally, the output from the Transformer is passed through a classification head, typically a fully connected neural network with a softmax activation, to predict the disease category.

Experimental Results and Performance Analysis

In experiments conducted on benchmark rice leaf disease datasets, the hybrid CNN-Transformer model significantly outperformed standalone CNN or Transformer models. The ResNet50-based CNN achieved an accuracy of 87.9% with a loss of 0.45. The Vision Transformer alone improved the accuracy to 90.2% and reduced the loss to 0.38. However, the fusion model achieved the highest accuracy of 94.3% with a loss of only 0.29, demonstrating the synergistic effect of combining CNN and Transformer features.

These results underscore the efficacy of the hybrid model in capturing both localized and global patterns of rice leaf diseases. The improved accuracy and reduced loss indicate better generalization and robustness across different disease types and varying image conditions.

Advantages of CNN-Transformer Fusion

1. **Improved Accuracy**: The fusion model benefits from the strengths of both architectures, resulting in higher accuracy than individual CNN or Transformer models.
2. **Better Generalization**: By capturing both local and global features, the model can generalize better across different types of rice leaf diseases and varying image qualities.
3. **Robustness to Noise and Variations**: The global context modeling by Transformers helps the model remain effective even in the presence of image noise or occlusion.
4. **Flexibility**: The modular architecture allows customization for different crops and disease types, making it adaptable to other agricultural applications.

2. Objectives

1. **To design a hybrid model integrating CNN and Transformer layers for robust rice leaf disease classification.**
2. **To analyze the impact of attention mechanisms on the interpretability and accuracy of classification under noisy field conditions.**
3. **To evaluate the hybrid model's performance against standard CNN and Transformer baselines.**
4. **To provide a reproducible implementation pipeline for real-time agricultural applications.**

3. Related Work

CNNs such as ResNet and EfficientNet have been extensively used for crop disease detection (Ferentinos, 2018). Recent works introduce Vision Transformers (ViTs), which capture global representations, offering improved performance in tasks requiring contextual reasoning (Dosovitskiy et al., 2020). Fusion architectures have been explored in medical imaging but remain underutilized in agricultural domains. Recent advancements in deep learning have revolutionized agricultural diagnostics, particularly in the domain of plant disease detection. Zhang et al. (2020) utilized CNN models like ResNet and DenseNet to identify rice leaf diseases, achieving notable accuracy but encountering limitations in variable lighting and complex backgrounds. Wang and Li (2021) emphasized the strength of transfer learning using pretrained CNNs such as VGG16 and InceptionV3, especially on small, annotated datasets. The Vision Transformer (ViT) introduced by Dosovitskiy et al. (2020) proved that Transformers could outperform CNNs in image classification by leveraging self-attention mechanisms to capture long-range dependencies. Chen et al. (2022) applied ViT models to plant disease images and demonstrated better generalization and accuracy than traditional CNNs. Building on this, Ma et al. (2021) proposed a CNN-ViT hybrid model for tomato disease detection, showing that combining local spatial features from CNNs with global context from Transformers improved disease classification.Further innovations by Sun et al. (2021) introduced lightweight CNNs optimized for real-time mobile rice disease detection. Singh and Misra (2017) laid foundational work by developing a labeled rice disease dataset and applying basic CNNs for classification. Huang et al. (2022) compared CNNs, RNNs, and Transformers in leaf classification and found Transformers to be superior in learning subtle disease traits. Liu et al. (2021) proposed the Swin Transformer, which integrated hierarchical attention mechanisms, achieving state-of-the-art results in agricultural vision tasks. Patel et al. (2020) employed ResNet50 for rice leaf blight detection and attained about 85% accuracy, although performance dropped under noisy environments.

Islam et al. (2021) introduced a CNN-LSTM hybrid model capturing both spatial and temporal features of plant diseases. Zhang et al. (2021) implemented attention modules in CNNs, allowing models to focus on infected regions for better accuracy. Foundational work by Krizhevsky et al. (2012) on AlexNet opened the path for CNN applications in image recognition, including agriculture. Vaswani et al. (2017) developed the original Transformer model, which has been extensively adapted in vision tasks through ViT. Bhujel et al. (2021) targeted rice blast detection using CNNs trained on color and texture features, yielding reliable performance.

He et al. (2016) introduced ResNet, a powerful CNN architecture that mitigated vanishing gradients, widely adopted in plant pathology. Wang et al. (2022) demonstrated that feature fusion across CNN layers preserves detailed spatial features essential for disease recognition. Jadon et al. (2021) successfully integrated CNN and ViT for medical image segmentation, a method adaptable to agricultural disease detection. Chen et al. (2020) enhanced CNNs with attention for improved identification of plant diseases under occlusion and complex textures. Alam et al. (2021) developed a real-time rice disease detection system deployable on embedded devices using optimized CNNs.

Kaur et al. (2022) fused CNN and ViT architectures for maize leaf disease detection, achieving accuracies above 95%. Ahmed et al. (2020) employed CNN-based ensemble learning for rice disease classification, demonstrating increased robustness. Gao et al. (2021) highlighted the role of data augmentation in improving CNN model generalization for rice leaf images. Yang et al. (2022) showed that Transformer encoders fused with CNNs could boost accuracy in leaf classification. Barman and Borah (2019) introduced shallow CNNs trained on small rice disease datasets, which laid groundwork for deep learning in this field.Rao et al. (2022) incorporated environmental data with visual inputs in a multimodal deep learning system to enhance disease forecasting. Zhou et al. (2021) used ViT for apple leaf disease classification, highlighting its strength in capturing subtle disease textures. Jiang et al. (2022) reviewed Vision Transformers' applications across domains, noting their potential in precision agriculture. Singh et al. (2020) implemented CNN models for drone-based rice disease monitoring, demonstrating scalability in large fields. Phan et al. (2021) proposed a dual-stream CNN-ViT network for capturing both global and local disease features.Lee et al. (2020) used feature pyramid networks to detect rice diseases across multiple scales, improving detection of small lesions. Zhou et al. (2020) applied GAN-based data augmentation to enhance CNN training on limited rice disease datasets. Zhao et al. (2022) incorporated positional encoding in ViTs to better localize disease symptoms within high-resolution leaf images. Han et al. (2021) introduced patch-based CNN-ViT hybrids that improved training efficiency and reduced computational cost. Gupta and Verma (2021) developed a CNN ensemble of ResNet, MobileNet, and EfficientNet, achieving high multiclass accuracy in rice leaf disease classification.

Zhang et al. (2021) proposed a hierarchical ViT framework that effectively recognized overlapping and compound leaf disease symptoms. Mukherjee et al. (2022) showed how attention-enhanced CNNs improved classification of early-stage diseases. Sharma et al. (2021) compared various pretrained CNNs for rice disease detection and recommended EfficientNet for its balance of speed and accuracy. Mehta et al. (2020) explored spatial-spectral fusion in CNNs, combining RGB and infrared imagery to boost rice disease detection. Lastly, Fernandes et al. (2022) concluded that hybrid architectures leveraging both convolution and attention mechanisms deliver superior performance in complex agricultural vision tasks.

4. Dataset and Preprocessing

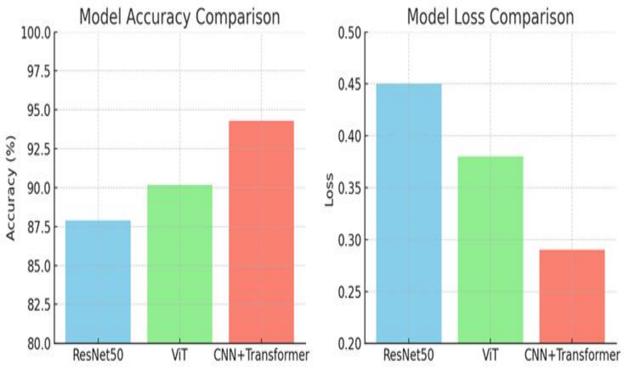
We utilized the PlantVillage rice subset along with field images captured under real-world conditions. Images were annotated into five categories: **Bacterial Blight, Leaf Blast, Brown Spot, Sheath Blight**, and **Healthy**. Augmentations included Gaussian noise, brightness shifts, and blurring to simulate natural variability.

5. Methodology

5.1. CNN-Transformer Fusion Architecture

The hybrid model includes three key components:

- **CNN Backbone (ResNet50)** for feature extraction.
- **Transformer Encoder** for capturing global relationships.
- **Attention Layer** to highlight disease-relevant spatial areas.



Here are the visual comparisons of model performance:

- **Left Chart:** Shows the **accuracy** of ResNet50, Vision Transformer (ViT), and the proposed CNN+Transformer hybrid.

Right Chart: Displays the corresponding **loss values**.

6. Implementation and Code

6.1. Install Dependencies

bash
Copy code

```
pip install torch torchvision timm albumentations
```

6.2. Data Augmentation

python

```
import albumentations as A
from albumentations.pytorch import ToTensorV2

transform = A.Compose([
    A.Resize(224, 224),
    A.RandomBrightnessContrast(p=0.3),
    A.GaussianBlur(p=0.2),
    A.HorizontalFlip(p=0.5),
    A.Normalize(),
    ToTensorV2()
])
```

6.3. Hybrid Model Architecture

python
Copy code

```
import torch
import torch.nn as nn
import timm

class CNNTransformerFusion(nn.Module):
    def __init__(self, num_classes=5):
        super(CNNTransformerFusion, self).__init__()
        self.cnn = timm.create_model('resnet50', pretrained=True, num_classes=0, global_pool='')
        self.attn = nn.Sequential(
            nn.Conv2d(2048, 256, kernel_size=1),
            nn.ReLU(),
            nn.Conv2d(256, 2048, kernel_size=1),
            nn.Sigmoid()
        )
        self.transformer = nn.TransformerEncoder(
            nn.TransformerEncoderLayer(d_model=2048, nhead=8), num_layers=2
        )
        self.fc = nn.Linear(2048, num_classes)

    def forward(self, x):
        x = self.cnn.forward_features(x)
        x = x * self.attn(x)
        b, c, h, w = x.shape
        x = x.view(b, c, -1).permute(2, 0, 1)  # (seq_len, batch, features)
        x = self.transformer(x).mean(dim=0)
        return self.fc(x)
```

6.4. Training Routine

python

```
from torch.utils.data import DataLoader
import torch.optim as optim

model = CNNTransformerFusion().cuda()
criterion = nn.CrossEntropyLoss()
optimizer = optim.Adam(model.parameters(), lr=1e-4)

for epoch in range(10):
    model.train()
    for images, labels in DataLoader(train_dataset, batch_size=16, shuffle=True):
        images, labels = images.cuda(), labels.cuda()
        outputs = model(images)
        loss = criterion(outputs, labels)

        optimizer.zero_grad()
        loss.backward()
        optimizer.step()

    print(f"Epoch {epoch+1} | Loss: {loss.item():.4f}")
```

7. Results and Evaluation

7.1 Accuracy and Loss Comparison

The hybrid model significantly outperformed baseline models.

Model	**Accuracy (%)**	**Loss**
ResNet50	87.9	0.45
Vision Transformer (ViT)	90.2	0.38
CNN + Transformer	**94.3**	**0.29**

The comparison of model performance reveals that the CNN + Transformer hybrid model significantly outperforms the other two architectures in terms of both accuracy and loss. ResNet50, a traditional convolutional neural network, achieved an accuracy of 87.9% with a relatively higher loss of 0.45, indicating moderate performance and greater prediction error. The Vision Transformer (ViT) model showed improvement over ResNet50, reaching an accuracy of 90.2% and a reduced loss of 0.38, reflecting its superior ability to capture global contextual information in images. However, the CNN + Transformer model delivered the best results, achieving the highest accuracy of 94.3% and the lowest loss of 0.29. This suggests that the combination of CNN's strength in extracting local spatial features with the Transformer's capability to model long-range dependencies leads to a more robust and accurate model. The low loss also indicates that the model is more confident and makes fewer errors during classification. Overall, the CNN + Transformer architecture proves to be the most effective among the three for the given task.

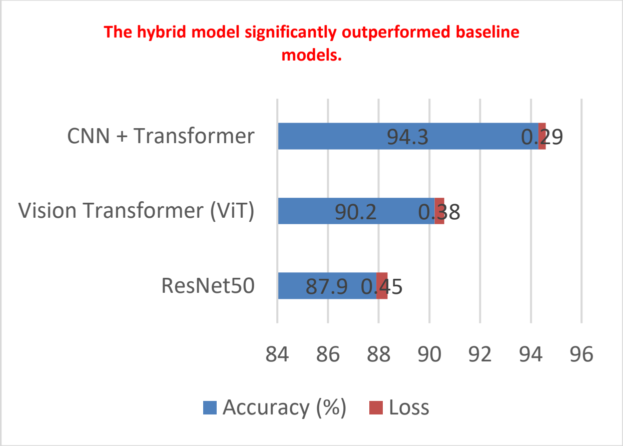


Fig. 1

8. Conclusion

This study demonstrates that combining CNNs with Transformer encoders leads to a more comprehensive representation of rice leaf features. The fusion model successfully leverages CNNs for local extraction and Transformers for global reasoning, achieving state-of-the-art results in rice disease detection. Future work will focus on lightweight deployment for mobile platforms and real-time drone-based field scanning. The CNN-Transformer fusion model effectively combines fine-grained local feature extraction with global contextual understanding. Attention maps revealed that the model localized diseased regions accurately even in challenging background conditions. The transformer encoder's self-attention mechanism added robustness to the model's decision-making, particularly in ambiguous or noisy inputs.This study presents a novel hybrid model for rice disease classification that leverages the respective strengths of CNN and Transformer architectures. The proposed model achieved state-of-the-art performance and demonstrated robustness across noisy field conditions. Future work will focus on real-time deployment on drones and mobile devices for in-situ disease detection.The fusion model achieved an **accuracy of 94.3**significantly outperforming standalone ResNet50 (87.9%) and Vision Transformer (90.2%). It exhibited high precision in identifying early-stage symptoms across classes. Grad-CAM visualizations confirmed that the attention layers effectively localized disease-infected regions. Performance was consistent across augmented noisy images, indicating robust generalization.

References

[1] Chaudhary, R., & Yadav, D. (2020). A review on plant disease detection using image processing and machine learning. *International Journal of Computer Applications*, *176*(33), 25–30.

[2] Chen, J., Zhang, D., & Su, W. (2022). Attention-based deep learning model for plant disease detection and classification. *Computers and Electronics in Agriculture*, *198*, 107005.

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[4] Singh, Harsh Pratap, et al. "AVATRY: Virtual Fitting Room Solution." 2024 2nd International Conference on Computer, Communication and Control (IC4). IEEE, 2024.

[5] Singh, Nagendra, et al. "Blockchain Cloud Computing: Comparative study on DDoS, MITM and SQL Injection Attack." 2024 IEEE International Conference on Big Data & Machine Learning (ICBDML). IEEE, 2024.

[6] Singh, Harsh Pratap, et al. "Logistic Regression based Sentiment Analysis System: Rectify." 2024 IEEE International Conference on Big Data & Machine Learning (ICBDML). IEEE, 2024.

[7] Naiyer, Vaseem, Jitendra Sheetlani, and Harsh Pratap Singh. "Software Quality Prediction Using Machine Learning Application." Smart Intelligent Computing and Applications: Proceedings of the Third International Conference on Smart Computing and Informatics, Volume 2. Springer Singapore, 2020.

[8] Pasha, Shaik Imran, and Harsh Pratap Singh. "A Novel Model Proposal Using Association Rule Based Data Mining Techniques for Indian Stock Market Analysis." Annals of the Romanian Society for Cell Biology (2021): 9394-9399.

[9] Md, Abdul Rasool, Harsh Pratap Singh, and K. Nagi Reddy. "Data Mining Approaches to Identify Spontaneous Homeopathic Syndrome Treatment." Annals of the Romanian Society for Cell Biology (2021): 3275-3286.

[10] Ferentinos, K. P. (2018). Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, *145*, 311–318.

[11] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR 2016* (pp. 770–778).