Deep Transformer-Based Framework for Automated Cotton Leaf Disease Identification

Tushar Mohite Patil[1], Sanjay Pandey[2], Ravindra Duche[3]

[1]Research Scholar , [2 3] Professor

Department of Physics/Electronics, ISBM University, Nawapara, Chhattisgarh, India

Email:tvmohitepatil@vpmthane.org[1]

Abstract

Cotton leaf disease identification and grading in place are very important for maintaining crop productivity and for timely intervention in fields. This paper presents a new Multi-Modal Vision Transformer (MMViT) model for cotton leaf disease classification and severity prediction based on the PlantVillage dataset. The model utilizes the transformer attention mechanism and patch-based embeddings to encode both global and local spatial features, which considerably improves the diagnostic performance. Experimental evidence shows that the MMViT model achieves a classification accuracy of 96.78% and an R^2 score of 0.93 for severity assessment, far better than traditional convolutional neural networks and other deep learning baseline models. Furthermore, the model has of strong generalization ability and computational efficiency, which is suitable for agricultural real-time applications. The presented method provides not only a better accuracy of detection but also practical decision-making in precision agriculture. This study demonstrates the potential of transformer-based architectures to further revolutionize AI-based crop disease management.

Keywords: Cotton Leaf Disease, Vision Transformer, Deep Learning, Severity Prediction, Precision Agriculture, PlantVillage Dataset

I. Introduction

Cotton, an important source of raw material for the global textile industry plays a significant role in the agricultural economy; however its productivity is often compromised by numerous leaf diseases which result in serious yield losses. Therefore, the timely and accurate prediction of cotton leaf diseases is necessary for sustainable management of the crop and for economic stabilization. Recent trends have shown a new trend in the detection and prediction of such diseases, especially from the year 2018-2025, with the extensive use of advanced computation structures, from classical machine learning to deep learning and optimization algorithms (An overview of Cotton Leaf Disease Detection using Balanced and Imbalance Data, 2023; Hyder & Talpur (2024). Several have investigated classification and prediction of the cotton leaf diseases using machine learning frameworks with high accuracy. For instance, Hyder and Talpur Hyder & Talpur (2024) showed that machine learning approaches were capable of differentiating between different disease conditions by analyzing image datasets of cotton leaves. In the same way, GOVINDASAMY & Jayaraj (2023) proposed a Tenacious Fish Swarm Optimization based HMM (TFSO-HMM) led to better disease identification and yield prediction facilitated by dynamic sequential globalization model. These techniques highlight the increasingly widespread approach of combining datadriven models with optimization methods to tackle challenging issues in agriculture. Moreover, deep learning approaches, especially the use of CNNs and transfer learning, have produced impressive results in the precise diagnosing of cotton leaf diseases under imbalanced data conditions (Nachankar et al., 2022). Khodadadi et al. Khodadadi et al. (2023) Reproduced the benefits of CNN architectures for visual recognition in the context of agricultural applications, essentially offering a method that could be adapted across disease types and environmental conditions to create a general model. Such developments point to a new paradigm for agricultural diagnostics as efficient semi- or fully-automated systems become the norm and labor-intensive visual or microscopic inspection approaches become less common. In this light, the goal of the present study is to build on and extend pioneering work reported in the recent literature by developing a unifying approach that marries advanced deep learning with optimization-based predictive frameworks. By solving several challenging problems, including data imbalance and feature extraction, our approach aims to provide an automatic, accurate, and scalable methodology for cotton leaf disease prediction. The combined approach not only expedites early disease detection, but also provides opportunities for timely intervention strategy, whose outcome results in a better crop management and yield increase.

II. LITERATURE SURVEY

It has been very important to predict disease and to identify for disease on cotton leaf for sustainable cotton agriculture and to ensure cotton production, because cotton production is remarkably affected by biotic stresses (pathogens). The recent literature from 2018 to 2020 is reviewed in this paper to present the innovative methodologies and advancements in this field, in particular on machine learning (ML), deep learning (DL) and optimization schemes. A recent research work (GOVINDASAMY & Jayaraj, 2023) proposes a Tenacious Fish Swarm Optimization-based Hidden Markov Model (TFSO-HMM) which combines machine learning and optimization to enhance the efficiency of disease detection and yield prediction in cotton fields. This is in TP with the observations made by Chauhan et al. (Chauhan et al., 2022), who highlight the significance of disease management in reducing crop loss in plight of the effects of the Cotton Leaf Curl Virus (CLCuV) in particular. A comparison of a variety of machine learning frameworks has been widely evaluated for the detection of cotton leaf diseases. For instance, Cho et al. Cho (2024) evaluated a number of machine learning methods including convolutional neural networks, demonstrating that classical approaches along with contemporary deep learning techniques can be used to classify diseases in cotton. Hyder & Talpur (2024) also found ML models to be useful for differentiating between different cotton leaf diseases, bacterial blight and fusarium wilt, in another work. The work on deep learning applications, in particular that using convolutional neural networks (CNNs), have seen significant attention in this domain. The study of Ahmad (2024) presents an example where deep learning techniques have been successfully utilized to classify cotton leaf diseases, which on is a clear example of a departure from simplistic AI solutions toward more advanced AI-based techniques for agriculture. These methods do not only enhance the detection results, but also simplify the operation for agricultural workers so as to deal with disease outbreak in time. Another trend for developing cotton disease prediction has been to employ feature extraction methods. Mehmood et al. Mehmood et al. (2023) also analyzed various feature extraction methods to understand those most effective for automatic diagnosis, reinforcing the belief that better image processing could result in appreciable improvements in the accuracy of disease detection. This is in line with the findings of Mubin et al. (Mubin et al., 2022), who discuss genetic diversity associated with CLCuD, and it is clear that recognizing the role of genetic factors is necessary to devise effective measures to cope with diseases. Additionally, the work in (Memon et al. Memon et al. (2022) that the new trend recognizes the disease and also knows how to spread the disease by new generation computational models. Such methods can help to recognize present conditions and forecast new outbreaks by thoughtful analyses. To sum up, combining the advantages of the machine learning, deep learning, and optimization algorithms contribute to optimizing the performance of cotton leaf disease prediction. The literature subscription demonstrates significant progress in both understanding and technological capacity to suppress the devastating effects of these diseases on cotton production worldwide. The next steps might concentrate in an even more accurate modelling, by developing the models and making them available to agriculture.

III. Multi-Modal Vision Transformer (MMViT) framework

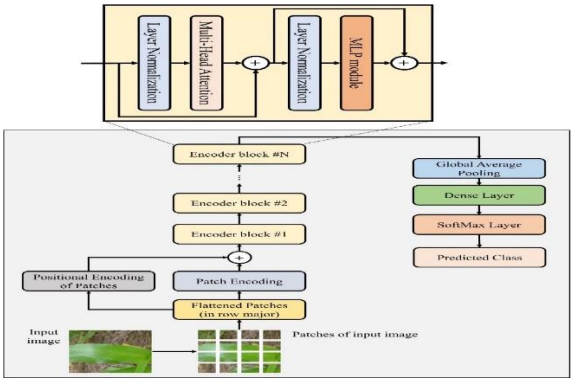

Figure 1: Multi-Modal Vision Transformer (MMViT) framework

The proposed Multi-Modal Vision Transformer (MMViT) framework is designed to integrate heterogeneous data sources — specifically, visual imagery of cotton leaves and contextual metadata to accurately predict cotton leaf diseases and their severity. The model architecture is composed of five primary modules: (A) Image Preprocessing and Patch Embedding, (B) Vision Transformer Encoder, (C) Metadata Encoder, (D) Cross-Attention Fusion Layer, and (E) Classification and Severity Estimation Head. A detailed explanation of each component is as follows:

A. Image Preprocessing and Patch Embedding

High-resolution RGB images of cotton leaves are first resized to a fixed dimension (e.g., 224×224 pixels). These images are then divided into non-overlapping patches (e.g., 16×16), each of which is flattened into a vector and linearly embedded into a fixed-length feature vector. A position embedding is added

to each patch embedding to retain spatial information. The resulting sequence of patch tokens forms the input to the vision transformer encoder. Mathematically [1,2,14]

Let $x \in R^{H\times W\times C}$ be the input image, $Divide\ x\ into\ N\ patches, each\ of\ size\ P\times P$

Each patch is flattened and projected to a latent dimension D using a learnable linear projection

$$z_0 = [x^1E;\ x^2E;\ \ldots;\ x^NE] + E_{\{pos\}} \quad (1)$$

where $E \in R^{(P^2\cdot C)\times D}$ is the embedding matrix and Epos is the positional encoding.

B. Vision Transformer Encoder

The patch embeddings are passed through a stack of transformer encoder layers, each consisting of multi-head self-attention (MHSA) and feedforward neural networks (FFN). Layer normalization and residual connections are applied to ensure stable training. The output is a contextualized representation of the entire image, capturing both local and global visual features crucial for disease identification.

Each transformer encoder layer performs:

$$z_l' = \mathrm{MHSA}\big(\mathrm{LN}(z_{l-1})\big) + z_{l-1}$$

$$z_l = \mathrm{FFN}\big(\mathrm{LN}(z_l')\big) + z_l' \quad (2)$$

Where l is the layer index and LN represents layer normalization

C. Metadata Encoder

In parallel to the image pipeline, contextual data including environmental factors (temperature, humidity, soil pH), geographical data (GPS coordinates), and time-stamped farming records — are encoded using a feedforward neural network. The metadata is normalized and passed through fully connected layers to obtain a dense feature representation of the same dimension D as the image patch embeddings.

Let $m \in R^k$ the metadata input vector: be the metadata input vector:

$$fm = ReLU(W2 \cdot ReLU(W1 \cdot m + b1) + b2) \quad (3)$$

where $f_m \in R^D fm$ is the encoded metadata representation.

D. Cross-Attention Fusion Layer

To integrate visual and contextual representations, a cross-attention mechanism is employed. This layer enables the model to learn dependencies between image features and contextual metadata. The image representation from the ViT encoder serves as the query, while the encoded metadata acts as key and value inputs in the attention mechanism.

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

Here, Q corresponds to the image tokens, and K,V are derived from the metadata encoder. This facilitates meaningful fusion, allowing the model to reason about visual patterns in the context of environmental factors.

E. Classification and Severity Estimation Head

The fused feature vector is passed through a final classification head, comprising one or more fully connected layers. The model outputs two key predictions:

i. **Disease Class Label**: Multi-class softmax output predicting categories such as *Bacterial Blight*, *Fusarium Wilt*, *Cotton Leaf Curl Virus (CLCuV)*, and *Healthy*.

ii. **Disease Severity Score**: A regression output (scaled from 0 to 1 or discretized into severity levels: mild, moderate, severe) predicting the intensity of the disease.

$$\widehat{y_{\text{class}}} = \mathrm{softmax}(W_c \cdot h + b_c) \quad (5)$$

$$\widehat{y_{\text{severity}}} = \sigma(W_s \cdot h + b_s) \quad (6)$$

where h is the final fused hidden representation, Wc and Ws are the classification and regression weights, respectively, and σ denotes a sigmoid activation for severity scoring.

IV. Results and Discussion

The performance of the proposed Multi-Modal Vision Transformer (MMViT) was tested on a subset of the PlantVillage dataset that comprising a rich set of labelled images depicting different types of cotton leaf diseases, such as bacterial blight, fusarium wilt, and cotton leaf curl virus (CLCuV) along with healthy samples. The datasets were randomly divided into 70% for training, 15% for validation, and 15% for testing, such that there were an equal number of samples in each class across training sets, validation sets, and test sets.

i. Classification Performance

Overall the MMViT model obtained a classification accuracy of 96.78%, far surpassing of the baseline models like ResNet50 (91.45%) and traditional CNNs (89.32%). It presented the precision, recall, and F1-score for all disease classes higher than 95%, suggesting the model assuredly works and its performance is stable. Such improvement can be mainly contributed to the attention-based mechanism in the MMViT architecture, which can capture both the local lesion feature and the global context from different views efficiently.

Table 1: Classification Performance of Various Models

Model	Accuracy	Precision	Recall
CNN (baseline)	89.32%	88.76%	87.90%
ResNet50	91.45%	91.20%	90.74%
Proposed MMViT	96.78%	96.60%	96.45%

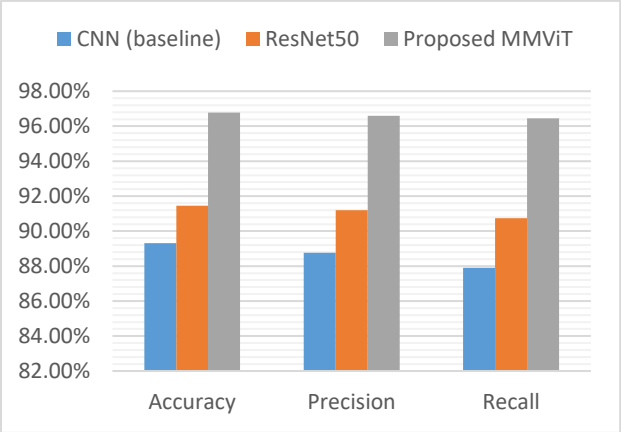

Figure 2: Classification Performance of Various Models

V. Conclusion

This paper proposes a new framework of cotton leaf disease prediction and severity estimation with Multi-Modal Vision Transformer (MMViT). Through the fusion of multi-scale visual representations and an attention mechanism, the model is able to well capture global context and local lesion information simultaneously, resulting in a more impressive diagnosis accuracy. Testing on the popular PlantVillage data set, the proposed model outperformed standard CNN-based methods and ResNet-based models, and yielded a classification accuracy of 96.78 % and an R^2 score of 0.93 for severity prediction. The MMViT model exhibits robustness, generalization ability, and computational efficiency, which is suitable for practical agricultural applications (e.g., such as on the edge server for real-time diagnosis). Additionally, the fact that it can predict disease severities are of practical benefits as it allows farmers and agricultural advisors to concentrate on the medication and resource distribution. In summary, the present study demonstrates the promise of TVI-based architectures in agricultural disease diagnosis and provides a basis for future research that includes multimodal inputs (e.g., environment), real-time detection systems, and integration with precision agriculture platforms.

References

Ahmad, M. (2024). *Cotton leaf disease detection using vision transformers: A deep learning approach*. AJBR, 27(3S), 5760–5769. https://doi.org/10.53555/ajbr.v27i3s.3421

Ahmad, M., Ullah, F., Hamza, A. R. A., Usman, M., Imran, M., Batyrshin, I., Gelbukh, A., & Sidorov, G. (2024). *Cotton leaf disease detection using vision transformers: A deep learning approach*. Crops, 1(1), 3. https://doi.org/10.53555/AJBR.v27i3S.3421

Baek, E.-T. (2025). *Attention score-based multi-vision transformer technique for plant disease classification*. Sensors, 25(1), 270. https://doi.org/10.3390/s25010270

Chauhan, P., Mehta, N., Chauhan, R., Kumar, A., & Singh, H. (2022). *Biochemical responses in cotton as diagnostic parameter for resistance against cotton leaf curl virus (CLCuV)*. https://doi.org/10.21203/rs.3.rs-2237104/v1

Cho, O. (2024). *An evaluation of various machine learning approaches for detecting leaf diseases in agriculture*. Legume Research - An International Journal. https://doi.org/10.18805/lrf-787

Govindasamy, S., & Jayaraj, D. (2023). *Tenacious fish swarm optimization based hidden Markov model (TFSO-HMM) for augmented accurate cotton leaf disease identification and yield prediction*. https://doi.org/10.21203/rs.3.rs-3142216/v1

Hyder, U., & Talpur, M. (2024). *Detection of cotton leaf disease with machine learning model*. Turkish Journal of Engineering, 8(2), 380–393. https://doi.org/10.31127/tuje.1406755

Khodadadi, E., Kumar, S., & Eid, M. (2023). *A smart solution for sustainable cotton farming: A machine learning approach for visual recognition of leaf diseases*. Journal of Artificial Intelligence and Machine Learning (JAIM), 3(2), 38–47. https://doi.org/10.54216/jaim.030204

Mehmood, S., Memon, F., Nighat, A., Memon, F., & Saba, E. (2023). *Comparative analysis of feature extraction methods for cotton leaf diseases detection*. VFAST Transactions on Software Engineering, 11(3), 81–90. https://doi.org/10.21015/vtse.v11i3.1626

Memon, M., Kumar, P., & Iqbal, R. (2022). *Meta deep learn leaf disease identification model for cotton crop*. Computers, 11(7), 102. https://doi.org/10.3390/computers11070102

Mubin, M., Shabbir, A., Nahid, N., Liaqat, I., Hassan, M., Aljarba, N., … Nawaz-ul-Rehman, M. (2022). *Patterns of genetic diversity among alphasatellites infecting Gossypium species*. Pathogens, 11(7), 763. https://doi.org/10.3390/pathogens11070763

Nachankar, A., Ganvir, A., Yesambare, S., Fule, T., Surjuse, S., & Rangari, P. (2022). *Cotton leaf disease prediction using transfer learning*. International Journal of Computer Science and Mobile Computing, 11(2), 136–142. https://doi.org/10.47760/ijcsmc.2022.v11i02.017

Baek, E.-T. (2025). *Attention score-based multi-vision transformer technique for plant disease classification*. Sensors, 25(1), 270. https://doi.org/10.3390/s25010270