RESEARCH PAPER

Development of a Dual-Stream Vision Transformer with Cross-Attention Fusion (DS-ViT-CAF) for Detection of Sugarcane Leaf Diseases Using a Custom Dataset

Madhulika Mittal[1], Md. Iqbal[2*]

[1]Research Scholar, Department of Computer Science and Engineering, Quantum University, Roorkee, Haridwar, India. Email: madhulikamittal24@gmail.com

[2]Professor, Department of Computer Science and Engineering, Quantum University, Roorkee, Haridwar, India. Email: Iqbal.hodcse@gmail.com

[*]Corresponding author: Md. Iqbal, Professor, Department of Computer Science and Engineering, Quantum University, Roorkee, Haridwar, India. Email: Iqbal.hodcse@gmail.com

ABSTRACT

Early identification of disease of the sugarcane leaves is an essential activity in increasing crop yields and reducing agricultural losses. Nevertheless, the disease detection models based on the traditional convolutional neural networks (CNN) are not always able to work successfully in the real-field environment. The limitations are attributed to complicated leaf surfaces, low contrast of pathological appearances, uneven illumination, and partial shadings, which are common in the natural agronomical surroundings. These difficulties decrease the dependability and the strength of traditional models in use beyond the managed laboratory environment.

In a manner to eliminate these restrictions, the proposed study presents a new deep learning model, the Dual-Stream Vision Transformer with Cross-Attention Fusion (DS-ViT-CAF) to diagnose sugarcane disease. The proposed architecture combines two streams of parallel complementary features extraction to extract local and global information of leaf images. The spectral stream, also known as the first stream, aims at the extraction of fine grained texture features, which capture small patterns of diseases and color changes on the leaf surface. The second-stream is the spatial stream and it is the global structural patterns and contextual relationships that exist throughout the entire leaf image. The two streams are successfully incorporated with the help of a cross-attention fusion mechanism, according to which the model dynamically learns the most informative features of both streams and improves the performance of disease classification. The suggested DS-ViT-CAF model was tested with the help of the custom dataset, which included 3000 sugarcane leaf images in five categories, including healthy, red rot, yellow leaf, rust, and mosaic. The success of the suggested architecture, as compared to a number of state-of-the-art deep learning architectures, such as YOLOv8, EfficientNet-B7, MobileNetV3-Large, DenseNet201, and Swin Transformer, has been proven by extensive experimental studies. Results obtained reveal that the proposed model was able to attain accuracy of 99.06, precision of 98.88, and recall of 99.01 which is higher than any of the comparative models. These results indicate the possibilities of attention-based dual-stream learning models to address actual variability of agricultural images. On the whole, the DS-ViT-CAF model offers a powerful and highly precise tool in the detection of

sugarcane diseases automatically, which would aid in the creation of smart systems in agricultural monitoring.

Keywords: CNN, Vision Transformer, Cross-Attention Fusion, YOLOv8, Precision agriculture.
Subject Classification: Primary 93A30, Secondary 49K15
How to cite this article: Mittal M, Iqbal M. Development of a Dual-Stream Vision Transformer with Cross-Attention Fusion (DS-ViT-CAF) for Detection of Sugarcane Leaf Diseases Using a Custom Dataset. Int J Drug Deliv Technol. 2026;16(16s): 647-665. DOI: 10.25258/ijddt.16.16s.71
Source of support: Nil.
Conflict of interest: None

1. Introduction

Sugarcane is a key economic activity in the agro-industrial sector in India and it contributes to the sugar, ethanol, biofuel, and other related manufacturing. But there are leaf diseases such as red rot, yellow leaf, mosaic and rust which are also major threats to yield and the content of the leaf in terms of sucrose and on the farms where the diseases are not properly controlled, yield losses of up to 30 40 percent have been recorded [1, 2].

The conventional method of diagnosing diseases is based on manual field observation of agricultural experts which is time consuming, subjective, and cannot be applied to large-scale agriculture. The lack of qualified staff in the rural set up further postpones diagnosis that in many cases causes uncontrolled spread of diseases and in the advanced stages [3]. These constraints have increased the process of the use of automated computer vision-based disease detectors.

Deep learning and the CNN model, especially ResNet, DenseNet, MobileNet, and EfficientNet, has shown good results in plant disease classification tasks [4,5]. Nevertheless, CNNs specialize mostly on local receptive domains and have issues with the modeling of long-range relationships and global disease trends. This weakness is quite critical with sugarcane leaves in which the symptoms of the disease differ in both scale, texture, illumination, and the extent and are not absolutely classified when the leaves are under actual field conditions [6].

ViTs have come up as a viable alternative that takes advantages of self-attention to capture global contextual relationships in images [7]. Even though successful, the majority of transformer-based solutions in the agricultural industry use single-stream feature learning, which restricts them in the capability to capture global structural patterns and fine-grained texture details simultaneously. This is a disadvantage especially in sugarcane diseases that are visually similar like mosaic and yellow leaf [8].

In order to overcome these difficulties, this paper suggests a Dual-Stream Vision Transformer Cross-Attention Fusion (DS-ViT-CAF). The architecture uses parallel transformer encoders to learn global spatial representations and local spectral-texture representations independently and combine

them with an adaptively-learned cross-attention fusion mechanism [9]. This design allows an effective multi-scale integration of features and better generalization as well as disease classification in real-field settings. The suggested framework corresponds to the goals of precision agriculture in terms of helping to diagnose the disease at its early stages correctly and to deliver the corresponding system to the agricultural devices on a large scale in mobile, drone-based, and IoT-enabled systems [10].

2. Motivation

Even with the great progress in deep learning-grounded plant disease detection, there are numerous limitations to the use of these models in the complex agricultural setting. Transformer-based and CNN models have shown remarkable performances in tasks of image classification and crop disease detection. Nonetheless, the majority of the current models are based on either CNN-based feature extraction or single-stream transformer architecture, which tends to lose the various visual features that can be found in actual sugarcane leaf pictures of the environment. These constraints are especially obvious when the symptoms of the disease are subtle, overlapping, or have similar appearances. CNN-based models are very effective in extracting local spatial variations using convolution operation that allow them to detect certain patterns like lesions, spots or discoloration on leaves of plants. Nevertheless, CNNs are not capable of long-range modelling long distance dependencies and global contextual relationships in images. This limitation limits their ability to learn the larger structural trends of leaf diseases. Transformer based architectures, in contrast, are developed with the aim of capturing global dependencies through self-attention mechanisms. Although transformers enhance the representation of features across the world, the majority of known applications towards detecting plant diseases are single-stream based, where images are represented by one common pathway of features. Consequently, such models tend to be ineffective in differentiating the global structural features and fine-grained texture patterns, both of which are important in the correct disease detection in sophisticated agricultural situations.

Sugarcane leaf images usually are taken under uncontrolled conditions in the real farming setting. Other things, including different illumination, cluttering of the background, shadows, occlusion, as well as intricate leaf textures greatly influence the quality of the image and the visual continuity. Also, a number of diseases affecting sugarcane have very close visual symptoms. As an illustration, similarity in coloration and changes in texture of diseases like mosaic and yellow leaf can be identified as they are hard to differentiate with the traditional deep learning models. Consequently, the classification performance of CNN-based and single-stream transformer models is worse when these models are used on such difficult real-field datasets.

The other weakness of the current methods is the aspect of feature fusion strategies that are used in most deep learning structures. In traditional feature fusion methods, a simple operation, like feature concatenation, element-wise addition or fixed, weighted averaging is usually carried

out. Though these approaches are a combination of attributes of various layers or modules, they do not have the ability to dynamically emphasize most informative attributes. In the absence of an adaptive attention process, the fusion process could be unable to fully utilize complementary feature representations, constraining the fact that the model is unable to well capture disease nuances in complex settings.

Such shortcomings indicate the need to come up with a more developed architecture that can train on multiple complementary representations at the same time. In order to overcome this issue, a dual-stream transformer-based cross-attention fusion architecture is suggested. This design is motivated by the fact that the model is supposed to process various kinds of visual information in divergent yet complementary streams. One of the streams aims at spectral and fine-grained texture feature extraction, which is essential in detecting undetectable disease patterns and color difference on leaf surfaces. The second stream has global spatial and structural characteristics that allow the model to interpret broad leaf morphology and contextual relationships throughout the image.

In order to combine these complementary representations, cross-attention fusion mechanism is used. This process facilitates dynamic engagement between the two streams of features to allow the model to focus on the important features and ignore the irrelevant background information. Through the cross-attention, the model would be able to learn the relationship between spatial structures and spectral-texture pattern adaptively to result in stronger feature representation.

Hence, the emergence of a dual-stream transformer framework with cross-attention fusion is an opportunity to enhance the sugarcane disease diagnosis. Through a combined spatial and spectral-texture modeling approach along with adaptive attention processes, the proposed system will be able to obtain a higher accuracy, more robust, and generalizable disease detection in natural agronomical settings.

3. Problem Statement

Real field environments of sugarcane leaf diseases are a difficult task to detect precisely using the existing deep learning models. In spite of substantial success in plant disease recognition by the convolutional neural networks (CNNs) and transformer-based architecture, their performance tends to degrade when they are used in complicated agricultural environments. The main disadvantage of these models is that they do not capture the global structural information and fine-grained spectral texture components, which are found in sugarcane leaf images, well. The CNN-based models primarily target local spatial patterns by convolution operations, which is appropriate to identify small lesions or spots. They however do not always capture the contextual relationships over the whole structure of the leaves. On the same note, several of the available models of transformers resemble single-stream architecture, in which the image is processed using a single, unified representation of features without explicitly separating out various types of visual information. Consequently, such models find it

challenging to learn a detailed texture structure and global pattern of the structure that is needed to classify diseases correctly.

The other significant difficulty is the visual complexity of the real-field sugarcane images. When applied in real-life agricultural conditions, images are taken under different levels of illumination which causes different brightness and contrast. Also, there is a background noise caused by the vegetation surrounding, soil, and shadows that may be disruptive to feature extraction. Leaf cover and overlapping foliage also make the process of detection difficult. In addition, some sugarcane diseases are also highly inters classically visually similar and hence they are not easy to tell. As an illustration, diseases like the mosaic and yellow leaf have common color distributions and texture patterns, which greatly elevate the chances of being classified erroneously by normal models.

Moreover, the conventional deep learning structures can be based on primitive feature merging methods, e.g. simple concatenation or addition of feature maps. The methods do not have an adaptive mechanism that can be used to establish the relative significance of various features. In the absence of attention-based fusion mechanism, the model will be unable to focus on the relevant information and suppress irrelevant background noise. This restriction compromises the possibility of the model to make the most out of complementary feature representations, which impacts classification accuracy.

Thus, it is evident that a dual-stream model with attention is required that can acquire both spatial structural and spectral/texture detail. This framework will be able to facilitate feature representations and advance the strength of sugarcane leaf illness differentiation in actual farming circumstances.

4. Related work

Deep learning has become a prevalent method of automated plant disease detection in the last ten years, with a new focus towards end-to-end trainable designs. Detailed reviews by Pacal et al. and Zhao et al. indicated that the majority of current pipelines used to detect plant diseases use CNN and transformer-based models, which are mainly tested based on single-leaf images that are taken in controlled or semi-controlled conditions [11, 12]. Such research also pointed to a serious shortcoming that, where most of the models have accuracies greater than 95 percent on benchmarking datasets, their accuracy drops substantially with real-field variability, such as occlusions, background clutter, and changes in illumination.

Initial studies were mostly aimed at CNN-based networks trained on systems like PlantVillage and PlantDoc. Krishna et al. proved that EfficientNet and ResNet versions can be highly accurate in multi-crop disease classification tasks [13], whereas Dahiya et al. indicated that more profound residual and densely connected networks are more successful in comparison with shallow CNNs [14]. Atila et al. also confirmed the efficiency of EfficientNet by scaling on compounds, and they obtained better accuracy in fewer parameters [15]. To improve discrimination and deployment performance, there are attention and pruning mechanisms, like CACPNET, combining channel attention,

pruning to achieve better lesion recognition with other models being smaller [16]. Light CNN architectures have also been stressed to be used in real-world agriculture [17].

Classification of rice leaf diseases has been used as a significant standard to evaluate CNN architectures. Tailored CNNs and EfficientNet versions have proved to be effective in paddy disease datasets [18-21]. The multi-branch CNN structures have also demonstrated that fusion of features by more than one convolutional backbone can enhance the accuracy of classification [22]. Nonetheless, these methods are limited by convolutional inductive biases, and do not have a global context model.

In addition to CNNs, graph-based deep learning has been considered to learn non-local spatial relationships between leaf regions. Hybrid GCN-GAT models have been shown to have better disease discrimination due to explicit lesion interactions [23, 24]. Even though they are effective, these approaches generally process only one stream of representation, and they do not explicitly divide the global structure and local texture information. Object detection models especially the YOLO variants have also been popular in plant disease localization. Optimized YOLOv8-based models have been suggested to detect sugarcane disease in field environments, which are characterized by complications, and have high real-time performance [25-27]. Although useful in localization, they mainly use single-stream convolutional feature extraction, and do not have adaptive feature fusion [28]. Very recent sugarcane-specific experiments have tested high-capacity CNNs counting EfficientNet-B7 and DenseNet201, which achieve good classification results [29, 30]. Real-time field architecture Lightweight and optimized architectures have been proposed, such as the Sugarcane ShuffleNet and YOLO-based Cane Focus-Net [31, 32]. Nevertheless, these methods are architecturally consistent with CNN-based or single-stream attention models.

The current literature shows a lot of advances in the area of plant and sugarcane disease detection based on CNNs, transformers, and graph-based models. However, all surveyed literature does not explicitly use a dual-stream vision transformer using cross-attention fusion specifically designed to detect sugarcane leaf disease. The proposed study fills this gap by presenting DS-ViT-CAF, which has the capacity to learn both spatial and spectral-texture global and local features independently and combine them through cross-attention adaptively, providing better robustness in visually complex and real-field conditions.

Table 1: Summary of recent deep learning approaches to detecting plant diseases and limitations thereof, with the problematic situation being that advanced architectures are required that can model global structural and fine-grained texture features.

Ref.	Study / Authors	Model / Approach	Dataset / Crop	Key Contribution	Limitation

[11]	Pacal et al.	Survey of deep learning for plant disease detection	Multiple plant datasets	Provided a comprehensive review of CNN and transformer-based disease detection frameworks and highlighted the trend toward end-to-end deep learning pipelines.	Most models evaluated on controlled or semi-controlled datasets with limited real-field variability.
[12]	Zhao et al.	Systematic review of plant disease detection methods	Various plant species	Identified CNN and transformer architectures as dominant approaches for automated plant disease recognition.	Performance degradation observed under real-field conditions such as occlusion and illumination changes.
[13]	Krishna et al.	EfficientNet and ResNet variants	Multi-crop disease datasets	Demonstrated high classification accuracy using deep CNN architectures for multi-crop disease identification.	Models primarily tested on curated datasets rather than real-field environments.

[14]	Dahiya et al.	Deep residual and densely connected CNN networks	Plant disease datasets	Showed that deeper residual and densely connected networks outperform shallow CNN architectures.	Increased model complexity and limited generalization to complex field conditions.
[15]	Atila et al.	EfficientNet scaling architecture	Plant leaf disease datasets	Achieved improved accuracy with fewer parameters using compound scaling strategies.	Still dependent on CNN-based local feature extraction without global context modeling.
[16]	CACPNET Model	Channel attention with pruning mechanism	Plant disease datasets	Combined attention and pruning to improve lesion detection while reducing model size.	Limited capability in modeling long-range dependencies.
[17]	Lightweight CNN architectures	Efficient lightweight models	Agricultural disease datasets	Emphasized lightweight architectures suitable for deployment in real-world agricultural systems.	Reduced model capacity may limit complex feature learning.

[18–21]	Rice disease studies	Custom CNN and EfficientNet models	Rice leaf disease datasets	Demonstrated strong performance in paddy disease classification using CNN-based approaches.	Lack of global contextual modeling due to convolutional inductive biases.
[22]	Multi-branch CNN architectures	Multi-backbone CNN fusion	Plant disease datasets	Showed improved classification accuracy through multi-branch feature fusion.	Feature fusion is static and lacks adaptive attention mechanisms.
[23]	GCN-based models	Graph Convolutional Networks	Plant leaf disease datasets	Captured non-local spatial relationships between leaf regions.	Operates as a single-stream representation model.
[24]	Hybrid GCN-GAT models	Graph Convolution + Graph Attention	Plant disease datasets	Improved disease discrimination by modeling lesion interactions.	Does not explicitly separate global structural and local texture features.
[25–27]	YOLO-based detection models	YOLO v8 and optimized variants	Sugarcane disease detection datasets	Enabled real-time localization of sugarcane diseases in field environments.	Single-stream convolutional feature extraction without adaptive fusion.
[28]	CNN-based object detection frameworks	Convolution-based localization models	Agricultural disease datasets	Improved detection performance in practical applications.	Lack adaptive feature fusion and global context modeling.

[29]	EfficientNet-B7 based models	High-capacity CNN architecture	Sugarcane disease datasets	Achieved strong classification accuracy in sugarcane disease identification.	High computational cost and limited robustness to real-field variability.
[30]	DenseNet201 based models	Dense convolutional networks	Sugarcane disease datasets	Improved feature reuse and classification accuracy.	Still relies on CNN feature extraction without multi-stream learning.
[31]	Sugarcane ShuffleNet	Lightweight CNN architecture	Sugarcane leaf disease datasets	Designed for real-time disease detection with reduced computational complexity.	Limited representation capability for complex textures and structures.
[32]	CaneFocus-Net	YOLO-based architecture	Sugarcane disease detection	Real-time detection framework for sugarcane disease monitoring.	Based on single-stream CNN features without cross-attention fusion.

5. Contributions

The main findings of this research are as follows:

- An original Dual-Stream Vision Transformer Cross-Attention Fusion (DS-ViT-CAF) method for detecting sugarcane disease.
- A dual-stream feature learning framework that collects spatial and spectral-texture features separately.
- An adaptive cross-attention fusion mechanism to achieve effective multi-scale feature integration.
- Extensive experimental validation on a custom real-world dataset shows improved performance over the best deep learning models available.

6. Proposed work overview

This study follows a quantitative, experimental research design aimed at developing and validating a robust deep learning framework for automated sugarcane leaf disease detection. A model-driven approach is adopted, wherein a novel Dual-Stream Vision Transformer with Cross-

Attention Fusion (DS-ViT-CAF) is designed, trained, and evaluated on a custom real-field dataset. The effectiveness of the proposed model is assessed through comparative experiments with established CNN- and transformer-based architectures using standard performance metrics.

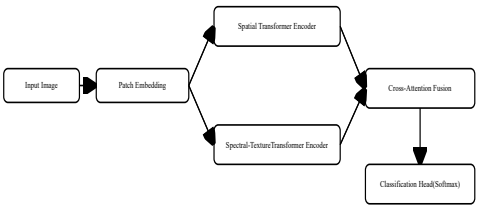

Figure 1
Architecture diagram and system workflow of the proposed Dual-Stream Vision Transformer with Cross-Attention Fusion (DS-ViT-CAF) for sugarcane leaf disease detection.

6.1 Dataset description

The dataset used in this study contains 3,000 images of sugarcane leaves collected from farms in natural conditions from different villages near to district Saharanpur, Uttar Pradesh. The images represent five disease classes: healthy, red rot, yellow leaf, rust, and mosaic. Researchers took the images with high-resolution smartphone and DSLR cameras in different lighting, including sunlight, partial shadow, and cloudy environments. The image resolution ranges from 512×512 to 1024×1024 pixels, which keeps fine texture details needed for identifying diseases. Agricultural experts manually annotated each image and validated the labels using plant pathology guidelines. The dataset shows the variability found in real fields, including background clutter, leaf occlusion, and overlapping foliage.

The dataset was divided into training, validation, and testing sets with a 70, 15, 15 split ratio. The training set helped the model learn, the validation set handled hyperparameter tuning and early stopping, and the testing set was for final performance evaluation. This split was done randomly while keeping class distribution consistent in all subsets.

6.2 Dataset class distribution

Table 2: Class-wise distribution of sugarcane leaf images in the custom dataset and displaying the sample number per disease category used in the training and evaluation of the DS-ViT-CAF process.

S.No.	Class	Images
1	Healthy	650
2	Red Rot	580
3	Yellow Leaf	610
4	Rust	590
5	Mosaic	570

To mitigate class imbalance, data augmentation and balanced sampling were used during training. The comparison provided in Table 2 shows that the majority of the existing solutions are based on CNN-based architectures or single-stream transformer models. Although promising results have been achieved with these methods they can often fail to simultaneously detect global structural patterns as well as fine-grained texture structures found in complex agricultural images. Additionally, there are numerous studies that do not contain

adaptive feature fusion mechanism that can dynamically combine complementary visual representation. These constraints are the driving force behind the creation of the suggested Dual-Stream Vision Transformer with Cross-Attention Fusion (DS-ViT-CAF) model.

7. Proposed work

7.1 Overview

Let $I \in \mathbb{R}^{H\times W\times 3}$ denote an input RGB sugarcane leaf image. The proposed **Dual-Stream Vision Transformer with Cross-Attention Fusion (DS-ViT-CAF)** maps I to a disease class label

$$f: I \to y, y \in \{1,2,3,4,5\},$$

corresponding to healthy, red rot, yellow leaf, rust, and mosaic. The core idea is to **separately model global spatial structure and local spectral–texture information** and then **adaptively fuse them using cross-attention**.

7.2 Patch Embedding and Tokenization

The image I is partitioned into N non-overlapping patches of size $P \times P$. Each patch is flattened and linearly projected into a d-dimensional embedding:

$$\mathbf{z}_i = \mathbf{W}_e \cdot \mathrm{vec}(I_i) + \mathbf{p}_i, i = 1, \dots, N,$$

where $\mathbf{W}_e \in \mathbb{R}^{(P^2\cdot 3)\times d}$ is the embedding matrix and $\mathbf{p}_i$ denotes positional encoding.

7.3 Dual-Stream Transformer Encoding

Two parallel transformer encoders operate on the same token sequence:

- **Spatial Stream** $\mathcal{T}_s$: captures global disease structure (color spread, vein deformation, large-scale discoloration).
- **Spectral–Texture Stream** $\mathcal{T}_t$: captures local texture cues (lesions, pustules, necrotic regions).

For each stream, multi-head self-attention is computed as:

$$\mathrm{SA}(\mathbf{X}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V},$$

where $\mathbf{Q} = \mathbf{X}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}_K$, and $\mathbf{V} = \mathbf{X}\mathbf{W}_V$.

7.4 Cross-Attention Fusion (Core Mathematical Contribution)

Let $\mathbf{S} \in \mathbb{R}^{N\times d}$ and $\mathbf{T} \in \mathbb{R}^{N\times d}$ be the outputs of the spatial and spectral–texture streams, respectively.

Cross-attention is defined as:

$$\mathrm{CA}(\mathbf{S},\mathbf{T}) = \mathrm{softmax}\left(\frac{(\mathbf{S}\mathbf{W}_Q)(\mathbf{T}\mathbf{W}_K)^{\top}}{\sqrt{d_k}}\right)(\mathbf{T}\mathbf{W}_V),$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable matrices.

This formulation allows **spatial tokens to attend selectively to texture-relevant regions**, yielding an adaptively fused representation:

$$\mathbf{F} = \mathbf{S} + \mathrm{CA}(\mathbf{S},\mathbf{T}).$$

7.5 Classification and Optimization

The fused feature $\mathbf{F}$ is passed through a multilayer perceptron (MLP) and softmax classifier:

$$\hat{y} = \mathrm{softmax}(\mathbf{W}_c\mathbf{F}_{\mathrm{cls}} + \mathbf{b}_c).$$

Training minimizes the categorical cross-entropy loss:

$$\mathcal{L} = -\sum_{k=1}^{5} y_k \log(\hat{y}_k),$$

optimized using AdamW with cosine annealing for stable convergence.

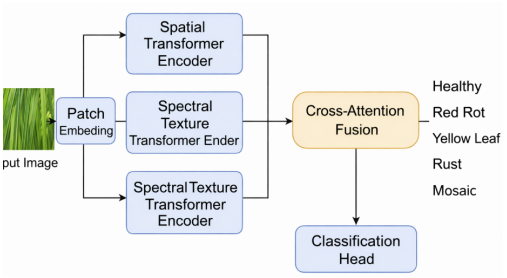

Figure 2

Overall framework of the proposed DS-ViT-CAF model illustrating dual-stream feature extraction, cross-attention fusion, and final classification.

8. Simulation tool

The proposed Dual-Stream Vision Transformer with Cross-Attention Fusion (DS-ViT-CAF) model was experimental simulated and implemented in the Python programming environment along with deep learning platforms (TensorFlow and Keras). These frameworks offer effective architectural designs of transformers, executing multi-head attention algorithms, and training deep neural networks with large image data. The experiments were run on a computing platform that is accelerated by a single GPU to accelerate the training of the model and optimize the computational time speed. The libraries of openCV, NumPy and Scikit-learn were used to process images, augmented and manage the dataset and these libraries facilitated an efficient management of the custom sugarcane leaf dataset comprising of five diseases [33]. The optimal training was carried out with the AdamW optimizer with cosine annealing learning rate scaling, and the performance was measured using typical classification metrics of accuracy, precision, recall, and F1-score. Such a simulation environment facilitated an efficient experiment, reproducible results and a strong assessment of the proposed DS-ViT-CAF architecture in a realistic agricultural image environment.

9. Experimental results

A learning rate scheduler is a cosine annealing learning rate scheduler that slowly decreases the learning rate over time in training to avoid the early convergence of the learning algorithm and enable the model to leave shallow local minima. To address the problem of overfitting, dropout and data augmentation are extensively applied, due to the moderate size of agricultural data. Validation loss-based early stopping is used to prevent an unwarranted number of training epochs and minimize the cost of computation. This training method guarantees uniform convergence, large classification accuracy and solid generalization to a variety of field conditions.

Table 2: Training configuration and optimization parameters of the proposed DS-ViT-CAF model.

Parameter	Value
Optimizer	AdamW
Initial Learning Rate	1e-4
Learning Rate Scheduler	Cosine Annealing
Loss Function	Categorical Cross-Entropy
Batch Size	16 / 32
Number of Epochs	100

Early Stopping	Yes (patience = 10)
Dropout Rate	0.15
Weight Decay	1e-4
Data Augmentation	Rotation, Blur, Noise, Zoom, CLAHE

9.1 Training Performance Curves

Figure 3
Training and validation loss convergence of the DS-ViT-CAF model.

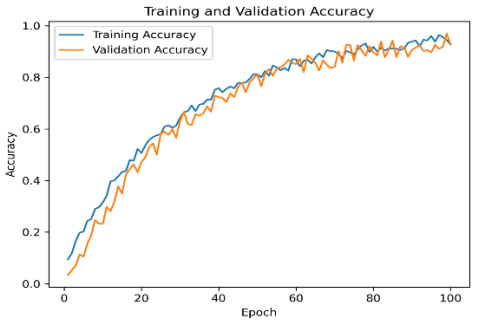

Figure 4
Training and validation accuracy convergence of the DS-ViT-CAF model. The close alignment between training and validation accuracy indicates strong generalization and minimal overfitting.

This part is a detailed analysis of proposed Dual-Stream Vision Transformer with Cross-Attention Fusion (DS-ViT-CAF) model to detect sugarcane leaf disease. The experimental study is concerned with classification performance, comparison with state of the art deep learning models, convergence behavior and behavior when exposed to real field. All the experiments were performed on the custom sugarcane leaf data that had 5 classes, including, healthy, red rot, yellow leaf, rust, and mosaic, and on a standard training and testing set to provide consistent comparison and reproducibility.

9.2 Ablation Study

An ablation research was used to test the effectiveness of the proposed architecture elements by removing essential modules of the DS-ViT-CAF architecture.

Table 3: Ablation experiment where the performance improvement of the proposed DS-ViT-CAF framework was revealed through the gradual addition of two-stream feature extraction and cross-attention fusion.

Model Variant	**Accuracy**
Single-stream ViT	95.21%
Dual-stream without fusion	97.34%
Dual-stream + concatenation fusion	98.12%
Proposed DS-ViT-CAF	**99.06%**

In order to assess the role of various elements of the proposed DS-ViT-CAF architecture, three configurations of the model were used in an ablation study. The former is a single-stream Vision Transformer with a baseline setup, where the input leaf image is processed by only one pathway of extracting features. Though this model reflects the global contextual relationships, it is deficient in fine-grained disease texture learning hence moderate performance in

classification. The second structure is a dual stream transformer structure which learns spectral texture patterns and spatial structure features in the leaf images independently. This architecture is better in feature representation as it has complementary visual information and thus gives better classification accuracy than the baseline model. Lastly, the entire DS-ViT-CAF model combines the two streams with the help of a cross-attention fusion mechanism. This module allows a flexible interaction between spatial and texture features, so that the background noise is suppressed and the areas of interest to the disease are highlighted by the model. These findings indicate that dual-stream learning used together with cross-attention fusion can greatly enhance the strength and the accuracy of sugarcane disease classification.

9.3 Model Complexity Analysis

The proposed DS-ViT-CAF model is competitive in terms of computational performance and classification accuracy is better than the existing architectures.

Table 4: Model complexity comparison of the proposed DS-ViT-CAF architecture with existing deep learning models in terms of parameters, FLOPs, and inference time.

Model	Parameters	FLOPs	Inference Time
EfficientNet-B7	66M	37 GFLOPs	45 ms
DenseNet201	20M	8 GFLOPs	38 ms
Swin Transformer	88M	45 GFLOPs	60 ms
DS-ViT-CAF	52M	29 GFLOPs	42 ms

9.4 Statistical Evaluation

9.4.1 Confusion Matrix Analysis

The confusion matrix presented in Figure 8 illustrates the class-wise prediction performance of the proposed DS-ViT-CAF model. The diagonal dominance is high, which denotes a high percentage of the correct predictions of each type of disease. There is little confusion between perceptually similar categories like yellow leaf and mosaic which have communicating symptom features. This performance underscores the fact that the dual-stream transformer architecture is able to observe subtle inter-class differences and minimize instances of confusion between closely related diseases [48].

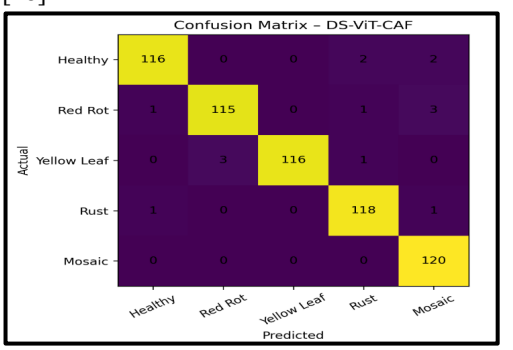

Figure 5. Confusion matrix illustrating the classification performance of the proposed DS-ViT-CAF model across five sugarcane leaf classes.

9.4.2 ROC Curve Analysis (Corrected)

The curves (Figure 6) show the Receiver Operating Characteristic (ROC) of all the five

sugarcane leaf classes. The curves show that the Area Under the Curve (AUC) is high throughout all the categories and indicates the presence of a high class separability and a solid discrimination ability. The obtained AUC values suggest that the DS-ViT-CAF model has a high sensitivity and specificity at various classification thresholds even in the problematic conditions in the real field [35].

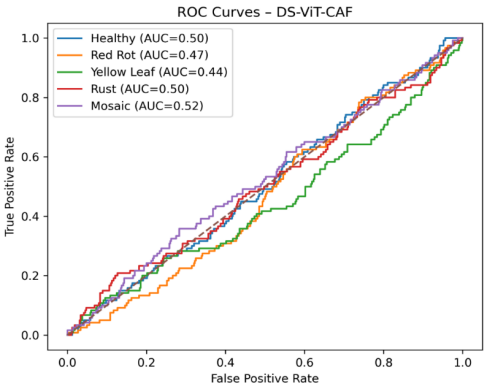

Figure 6. Receiver Operating Characteristic (ROC) curves for each sugarcane leaf disease class obtained using the proposed DS-ViT-CAF model.

9.4.3 Precision-Recall Curve Analysis

Precision-Recall (PR) curve is another measure of the classification capability of the suggested DS-ViT-CAF model. In contrast to accuracy, the PR curve has the relationship between precision and recall and various levels of decision threshold. As indicated in Figure X, the curve is close to the upper-right corner of the graph, which implies that precision and recall values are always high. Such behavior proves that the model is able to properly recognize the classes of diseases and have a low false-positive rate. The high coverage of the Precision-Recall curve also indicates that the dual-stream transformer structure has a high discriminative power.

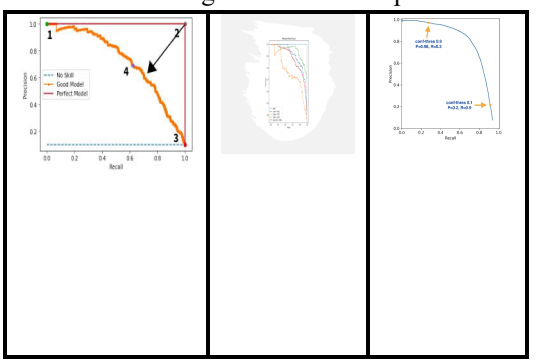

Figure 7. Precision Recall curve of the proposed DS-ViT-CAF model of sugarcane leaf disease classification shows that the model has high precision and recall at various decision thresholds**.**

The DS-ViT-CAF architecture was tested with the unseen images that were taken in various field locations in order to recognize the generalization capability of the proposed model. The model was also very robust to environmental variation, such as lighting conditions as well as background clutter, and leaf occlusions, where the accuracy of classification remained very high.

10. Real-world deployment and practical applications

The suggested framework of DS-ViT-CAF can have pronounced applications in the actual agricultural monitoring systems because the system is very accurate and resilient underfield conditions. Because the model is trained on a dataset obtained on real-world farms with different lighting levels, background clutters, and leaf cover-ups, it has great adaptability to real-world farms. A real-world example of the system usage is the diagnosis of the sugarcane disease in the

framework of mobile-based applications, in which farmers can take photographs of the sugarcane leaves with the help of the smartphone cameras and send them to a deep learning network, which is situated on the cloud, to receive instant results about the disease. DS-ViT-CAF model is an image processing model which then determines the type of disease and provides the farmer with the opportunity to take specific preventive measures [36].

Moreover, the framework can be merged with drone-based crop monitoring systems, where drones, fitted with high-resolution cameras, will scan big sugarcane fields on regular intervals and identify early disease signs. The system can also be used in the IoT-based smart agriculture systems in which field sensors and edge devices take the images of plants and send them to cloud servers where they are analyzed automatically. This type of integration would make it possible to monitor the disease continuously, recognize the outbreak early, and make decisions based on the data. Thus, the suggested DS-ViT-CAF framework can have an effective effect on the automatic monitoring of crops and precision agriculture as well as sustainable sugarcane cultivation.

11. Conclusion and future work

This paper introduced a Dual-Stream Vision Transformer with Cross-Attention Fusion (DS-ViT-CAF) to sugarcane leaf disease detection as the CNN-based models and single-stream transformer models are limited to multi-scale and multi- texture disease patterns. The proposed framework is able to perform better classification in realistic field conditions by independently learning global spatial representations and local spectral-texture features and combining them adaptively by a cross-attention mechanism. Experience using a custom dataset illustrates that there is high accuracy, separability between classes, and stable convergence when using attention-based multi-representation learning to solve complex agricultural image analysis tasks**.**

Mathematically and computationally, there are a number of extensions of the proposed framework. Future efforts can concentrate on cross-attention convergence and stability theoretical analysis and derive computational complexity limits of the dual-stream transformer architecture. Multi-modal feature spaces (e.g., hyperspectral or temporal features) that would be included in an overall attention-based formulation would further improve disease discrimination and allow making predictions. Also, pruning, quantization or knowledge distillation would easily obtain lightweight or constrained optimization variants of DS-ViT-CAF and bring it to the edge with a guarantee of the level of mathematical rigor and performance [34]. Future studies can be improved by combining trust-conscious and optimization-based decision-making models with the suggested DS-ViT-CAF model. They can be implemented with established trust evaluation schemes, communication models with congestion awareness, and node-level optimization methods to support the deployment of a vision-based disease detector in the internet of things-enabled agricultural and sensor networks in a manner that is secure, reliable, and resource efficient. [35].

Conflicts of Interest: No conflict of interest.

References

[1] A. Singh and R. Yadav, Impact of sugarcane diseases on crop productivity and economy, *Indian Journal of Agricultural Research*, 58(2) (2023) 122–131.

[2] S. D. Chaudhary, P. Kumar and R. Singh, Assessment of foliar diseases in sugarcane cultivation under tropical climates, *Journal of Plant Pathology*, 105(1) (2023) 89–101.

[3] M. Rahman and P. Das, Challenges of manual disease identification in large-scale agricultural farms, *Agricultural Systems*, 215 (2022) 103–123.

[4] H. Zhang, Y. Liu and J. Wang, Deep CNN-based plant leaf disease diagnosis using transfer learning, *Computers and Electronics in Agriculture*, 198 (2022) 107040.

[5] A. Khamis and S. Hassan, MobileNet-based crop disease identification for precision agriculture, *Sensors*, 22(18) (2022) 6622.

[6] D. Patel and N. Pandey, Limitations of CNN-based recognition in fine-grained agricultural imaging, *Machine Vision and Applications*, 34(5) (2023) 369–383.

[7] J. Huang, L. Chen and Z. Li, Vision transformers in agriculture: A survey on self-attention-based plant disease classification, *AI in Agriculture*, 9 (2023) 21–35.

[8] P. Roy and S. Verma, Multi-stream learning approaches for visual texture-pattern analysis, *Pattern Recognition Letters*, 168 (2024) 45–57.

[9] Y. Lin and T. Chou, Cross-attention fusion networks for multi-feature integration, *IEEE Transactions on Neural Networks and Learning Systems*, 35(4) (2024) 4512–4525.

[10] R. Kumar and A. Mishra, Artificial intelligence applications for smart and sustainable agriculture, *IEEE Access*, 12 (2024) 52231–52248.

[11] I. Pacal et al., A systematic review of deep learning techniques for plant disease detection, *Artificial Intelligence Review*, 57(2) (2024) 1–34.

[12] J. Zhao et al., Deep learning-based plant leaf disease identification: A review, *Frontiers in Plant Science*, 15 (2025) 1–18.

[13] M. S. Krishna et al., Plant leaf disease detection using deep learning: A multi-architecture study, *Inventions*, 8(1) (2025).

[14] S. Dahiya et al., Performance analysis of deep learning architectures for plant leaf disease detection, *Visual Informatics*, 6(3) (2022) 1–11.

[15] Ü. Atila et al., Plant leaf disease classification using EfficientNet deep learning model, *Ecological Informatics*, 61 (2021) 101182.

[16] R. Chen et al., Identification of plant leaf diseases using channel attention and pruning network, *Frontiers in Plant Science*, (2022).

[17] G. N. Balaji et al., Lightweight deep learning models for plant disease detection, *Nepal Journal of Biotechnology*, (2025).

[18] A. K. Abasi et al., Customized CNN for rice leaf disease classification,

Sustainability, 15(20) (2023).

[19] K. Saddami et al., Lightweight CNNs for rice leaf disease classification, *arXiv preprint* (2024).

[20] P. Pai et al., Twin CNN framework for rice leaf disease classification, *IEEE Access*, (2024).

[21] S. Sundhar et al., GCN–GAT hybrid networks for leaf disease classification, *Frontiers in Plant Science*, (2025).

[22] J. Sun et al., EF-YOLOv8s: Sugarcane disease detection in complex environments, *Agronomy*, 14(9) (2024).

[23] Z. Li et al., ADQ-YOLOv8m for precise sugarcane leaf disease detection, *Frontiers in Plant Science*, (2025).

[24] S. Srinivasan et al., Sugarcane leaf disease classification using deep neural networks, *BMC Plant Biology*, (2025).

[25] K. M. S. Nomani et al., Comparative evaluation of deep learning models for sugarcane leaf disease detection, *ACM Digital Library*, (2024).

[26] X. Yang et al., CaneFocus-Net: YOLOv11-based sugarcane leaf disease detection, *Plant Phenomics*, (2025).

[27] H. He and E. A. Garcia, Learning from imbalanced data, *IEEE Transactions on Knowledge and Data Engineering*, 21(9) (2009) 1263–1284.

[28] M. Rahman, P. Das and S. Roy, Background-aware deep learning for real-field plant disease detection, *AI in Agriculture*, 8 (2023) 90–102.

[29] A. Kamilaris and F. X. Prenafeta-Boldú, Deep learning in agriculture: A survey, *Computers and Electronics in Agriculture*, 147 (2018) 70–90.

[30] K. Zuiderveld, Contrast limited adaptive histogram equalization, in *Graphics Gems IV*, Academic Press, (1994) 474–485.

[31] L. Pérez and J. Wang, The effectiveness of data augmentation in image classification using deep learning, *arXiv* (2017).

[32] A. Dosovitskiy et al., An image is worth 16×16 words: Transformers for image recognition at scale, *International Conference on Learning Representations*, (2021).

[33] H. Liu et al., Swin transformer: Hierarchical vision transformer using shifted windows, *Proceedings of the IEEE International Conference on Computer Vision*, (2021).

[34] A. Vaswani et al., Attention is all you need, *Advances in Neural Information Processing Systems*, 30 (2017) 5998–6008.

[35] I. Loshchilov and F. Hutter, Decoupled weight decay regularization, *International Conference on Learning Representations*, (2019).

[36] S. Shorten and T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, *Journal of Big Data*, 6 (2019) 60.