Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00369-4

Original Article

Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions

Moïse K. Kasambya [1, 2]

1. *Department of Computer Science, University of the Philippines Diliman, Quezon, PHL*
2. *Business and Information Technology, Université de l'Assomption au Congo, Butembo, COD*

Received: September 18, 2026 | Review began: September 22, 2026 | Review ended: October 12, 2026 | Published: November 15, 2026

© **Copyright** 2026
This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Banana diseases cause significant crop losses worldwide and threaten food security in many tropical and subtropical regions. Accurate disease diagnosis remains challenging under real-world field conditions characterized by cluttered backgrounds, conditions far removed from laboratory-controlled environments. This paper presents Social-AgriNet, an explainable multimodal deep learning framework for robust banana disease detection from field-acquired imagery. The proposed architecture integrates a Vision Transformer (ViT-Base/16) backbone, an EfficientNet-B2 auxiliary branch, an image-derived metadata encoder, and a gated multimodal fusion mechanism. Evaluated on a geographically independent Healthy-only test set from Tanzania (5,628 images), Social-AgriNet achieves 94.42% ± 2.12% binary healthy recognition accuracy under the default training configuration and 96.58% ± 0.59% under the optimal configuration (omitting label smoothing). Because the Tanzanian test set contains no diseased samples, multi-class disease detection is reported separately on the Asian validation split (n = 936, all nine classes represented), where the model achieves 92.74% accuracy and macro-F1 = 0.9304. Multi-class field generalization to geographically distinct regions remains unproven. Ablation studies confirm that MixUp augmentation (+5.68 pp) and class-weighted sampling (+3.70 pp) are the critical regularization components, while label smoothing is detrimental on this dataset. Fusion gate analysis reveals a continuous modality-weighting mechanism, and Grad-CAM visualizations confirm alignment with agronomically meaningful lesion regions.

Categories: Computer Vision, Machine Learning (ML), Deep Learning

Keywords: banana disease detection, multimodal fusion, vision transformer, spatial attention, explainable ai, heterogeneous real-world imagery

Introduction

Banana (Musa spp.) is one of the most economically and nutritionally important crops worldwide, supporting the livelihoods of an estimated 400 million people across tropical Africa, Asia, and Latin America [1]. Despite its global importance, banana production remains critically vulnerable to a constellation of fungal, bacterial, and viral pathogens. Black Sigatoka (*Pseudocercospora fijiensis*), Panama Wilt (*Fusarium oxysporum* f. sp. cubense), Anthracnose (*Colletotrichum musae*), Cordana Leaf Spot (*Cordana musae*), Pestalotiopsis Leaf Spot (*Pestalotiopsis* spp.), and Nutrient Deficiency collectively cause losses exceeding USD 1.5 billion annually [2,3].

Traditional disease diagnosis often depends on trained agronomists, who are inaccessible to many smallholder farmers in low- and middle-income countries. Computer vision and deep learning have emerged as promising technologies for automating plant disease detection from leaf images [4-6]. However, most existing systems are trained on laboratory-

quality datasets such as PlantVillage, where images are captured under uniform lighting against standardized backgrounds, conditions far removed from the noisy, heterogeneous photographs captured in the field via smartphones and uploaded to agricultural community platforms [7,8].

Real-world agricultural imagery presents unique challenges: severe background clutter, variable illumination, blur, compression artifacts, and occlusion. Furthermore, purely visual models ignore the rich contextual signals encoded in image-derived quality metadata-brightness, blur score, saturation, contrast, aspect ratio, quality score, hue variance, and edge density which carry meaningful information about image quality and domain shift [9,10].

Despite recent advances in plant disease classification, most existing approaches remain limited to single-modality visual analysis and are evaluated primarily on laboratory-controlled datasets. Consequently, their generalization ability under real-world field conditions remains insufficient for practical agricultural deployment. Recent work has demonstrated the importance of explainability and robustness in agricultural artificial intelligence (AI) systems, with studies on high-confidence misclassification in deep learning highlighting the need for careful validation [11]. Furthermore, comparative studies of MobileNet, ResNet50, and EfficientNetB0 for banana leaf disease classification have established strong baselines for vision-based diagnosis [12].

To address these limitations, we propose Social-AgriNet, a multimodal deep learning framework specifically designed for robust banana disease detection under realistic, real-world data conditions. The primary contributions of this study are summarized as follows:

1. We propose Social-AgriNet, achieving 94.42% ± 2.12% binary healthy-recognition accuracy (96.58% ± 0.59% under optimal training) on a geographically independent African field test set consisting exclusively of Healthy leaf images. Multi-class disease detection performance is reported separately on the Asian validation split (92.74% accuracy, macro-F1 = 0.9304).

2. We introduce a hierarchical spatial attention mechanism integrating channel attention, spatial attention, and metadata-guided cross-modal attention.

3. We evaluate the framework on field-acquired datasets under multiple image degradations including blur, low-light, and JPEG compression.

4. We provide comprehensive ablation studies (backbone, attention, fusion, training strategy) demonstrating the critical importance of MixUp augmentation and class-weighted sampling, while revealing that label smoothing degrades performance.

5. We provide explainability analysis using Grad-CAM visualizations to support interpretable agricultural decision-making.

Related work

Deep Learning for Plant Disease Detection

The seminal work of Mohanty et al. [5] demonstrated that convolutional neural networks (CNNs) trained on the PlantVillage dataset could achieve over 99% accuracy under controlled conditions. This work catalyzed a wave of agricultural disease detection studies employing transfer learning from ImageNet-pretrained architectures including AlexNet, VGG-16, ResNet, and Inception [4,13,14]. Saleem et al. [6] surveyed 40 deep learning papers on plant disease detection and identified performance drops of 30-50 percentage points when models trained on PlantVillage were evaluated on field-collected photographs, highlighting a significant domain gap.

For banana-specific disease detection, Barbedo [15] explored the impact of image resolution and quality, finding significant performance degradation under field conditions. Paymode and Malode [16] applied MobileNet and VGG architectures to banana leaf disease classification, achieving 98% accuracy on clean datasets but noting limitations under real-world noise. Zheng et al. [17] proposed a lightweight convolutional approach for mobile deployment, while Narayanan et al. [18] investigated ensemble methods for banana disease classification across six categories.

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

Recent comparative studies have established strong baselines for banana disease classification. Kasambya et al. [12] conducted a comparative study of MobileNet, ResNet50, and EfficientNetB0 with Grad-CAM visualization and statistical validation, providing a foundation for explainable disease diagnosis. Delphin et al. [19] introduced PlantVillageVQA, a vision-language reinforcement learning framework for explainable crop monitoring, demonstrating the potential of multimodal reasoning in agricultural AI.

Vision Transformers in Agricultural Imaging

Following the introduction of the Vision Transformer (ViT) by Dosovitskiy et al. [20], transformer-based architectures have increasingly been explored for agricultural image analysis. ViTs leverage self-attention across patch embeddings to capture global spatial dependencies, complementing local texture features emphasized by CNNs. Liu et al. [21] explored the Swin Transformer for crop disease detection, finding that hierarchical attention representations reduced sensitivity to background clutter.

Multimodal Learning in Precision Agriculture

Recent advances have explored integrating visual features with environmental metadata to improve agricultural decision support. Kamilaris and Prenafeta-Boldú [22] reviewed deep learning applications in agriculture, highlighting the potential of geospatial and environmental data fusion. Ramcharan et al. [10] demonstrated that incorporating GPS metadata with plant image classifiers improved disease mapping accuracy in cassava. Tsouros et al. [23] reviewed UAV-based precision agriculture applications where remote sensing data were fused with ground-level imagery. Kasambya et al. [11] investigated high-confidence misclassification in deep learning-based systems, highlighting the importance of robust validation and the "paradox of certainty" in agricultural AI. This finding underscores the need for explainable and statistically validated models in plant disease detection.

Attention Mechanisms for Agricultural Image Analysis

Attention mechanisms have become a cornerstone of modern computer vision architectures. The Squeeze-and-Excitation Network (SENet) introduced channel-wise attention [24], while the Convolutional Block Attention Module (CBAM) extended this to spatial attention [25]. Both mechanisms have been adapted for plant disease detection to suppress uninformative background regions. Fu and Shi [26] proposed an attention-guided framework for leaf disease segmentation. Zeng and He [27] applied CBAM to rice disease detection, reporting significant improvements in precision. Cross-modal attention - where one modality guides attention in another - has gained traction in vision-language tasks [28] and is increasingly applied in remote sensing data fusion [29].

Explainable AI in Plant Pathology

The adoption of deep learning in agricultural decision support has raised concerns about model interpretability. Grad-CAM [30] and its variants provide class-discriminative visualizations highlighting influential image regions. Paymode and Malode [16] applied Grad-CAM to banana disease classification, confirming alignment between attention maps and visible lesion regions. Kasambya et al. [12] used Grad-CAM to validate explainability in banana disease classification, demonstrating that attention maps consistently localize disease-relevant regions.

Delphin et al. [19] extended explainability to vision-language models, showing that reinforcement learning with visual grounding can produce human-readable diagnostic explanations. LIME [31] and SHAP [32] have been used to explain individual predictions in agricultural machine learning pipelines. As AI-based diagnostic tools are deployed in field settings, explainability becomes essential for farmer trust and regulatory compliance [33].

Heterogeneous Real-World Agricultural Imaging

Disease reporting through mobile agricultural platforms presents both opportunities and challenges. Yang et al. [34] characterized the visual quality distribution of heterogeneous real-world plant images, noting that smartphone images exhibit significant blur, compression, and lighting heterogeneity. Singh et al. [8] introduced the PlantDoc dataset as a benchmark for evaluating model performance on real-world field images. BananaLSD, a banana leaf disease dataset collected from a banana farm at Bangabandhu Sheikh Mujibur Rahman Agricultural University and nearby farmer fields in

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

Gazipur, Bangladesh, provides a realistic testbed for field-based diagnostics [35]. Real-world data have also been used to train participatory disease surveillance systems in cassava [36] and rice [37]. Kasambya et al. [11] highlighted the risks of high-confidence misclassification in deep learning systems, emphasizing the need for robust evaluation protocols and uncertainty quantification in agricultural AI deployments.

Materials And Methods

Problem formulation

Let an input sample consist of a leaf image $x \in R3 \times H \times W$ and an associated metadata vector $m \in R8$ encoding image-derived quality features. The target output space consists of C = 9 categories. However, due to regional domain constraints in field data acquisition, the primary out-of-domain evaluation set (Tanzania) comprises a binary subset distribution (Healthy vs. Disease). Evaluation metrics are reported both globally and conditioned on class availability to prevent statistical artifact collapse.

$$\hat{y} = \arg\max_{c} f_c(x, m), \quad c \in \{0, 1, \ldots, 8\}$$

Dataset construction

Data Sources

Social-AgriNet is trained on a composite dataset drawn from four sources spanning controlled to uncontrolled imaging conditions, all restricted to genuine banana leaf imagery.

BananaLSD [35]: A banana leaf disease dataset from Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur, Bangladesh. Contains 937 images across four classes: Healthy (129), Black Sigatoka (473), Cordana Leaf Spot (162), and Pestalotiopsis Leaf Spot (173). Images exhibit realistic field-acquisition degradations.

Tanzania Banana Dataset (Test only): A field-collected dataset of 5,628 Healthy banana leaf images from Tanzania, used exclusively as the geographically independent test set (Africa). All 5,628 test images belong to the Healthy class, providing a clean out-of-distribution evaluation of the model's cross-continental generalization without introducing class-imbalance artifacts in evaluation metrics. Because the Tanzania test set contains no diseased samples, it cannot be used to evaluate multi-class disease detection. It evaluates only binary Healthy vs non-Healthy discrimination. Multi-class performance is therefore reported on the Asian validation split (936 images, all nine classes represented).

Nutrient Deficiency Dataset: 3,100 images capturing banana plant nutrient deficiency symptoms (boron, calcium, iron, magnesium, manganese, potassium, sulpfur, zinc: 2,150 images) plus Healthy leaves (950 images) under diverse field conditions.

Banana_and_Banana_Leaf Dataset: 2,375 images spanning Healthy (1,132), Black Sigatoka (126), Anthracnose (839), Pest Damage (224), and Fusarium Wilt (54). Adds training diversity.

Dataset deduplication and geographic splits

Dataset Deduplication and Splits

Deduplication removed 143 duplicate images (MD5 hashing), resulting in 11,897 unique images. The geographic split ensures true cross-domain evaluation:

(i) Training (Asia): 5,333 images from BananaLSD, Nutrient Deficiency, and Banana_and_Banana_Leaf

(ii) Validation (Asia): 936 images (15% stratified split from training data)

(iii) Test (Africa - Tanzania): 5,628 images, exclusively Healthy class

Post-deduplication class distribution

Table *1* reports the exact class counts per split after deduplication, addressing reproducibility requirements.

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

Class	Train (Asia)	Validation (Asia)	Test (Tanzania)
Healthy	1,690	298	5,628
Black Sigatoka	473	83	0
Cordana Leaf Spot	138	24	0
Pestalotiopsis	147	26	0
Fusarium Wilt	46	8	0
Anthracnose	713	126	0
Pest Damage	190	34	0
Nutrient Deficiency	1,828	322	0
TOTAL	5,333	936	5,628

TABLE 1: Post-deduplication class distribution per split. Tanzania test set comprises exclusively Healthy images.

This geographic split ensures that the model is evaluated on truly unseen data from a different continent, mimicking real-world deployment scenarios where models must generalize to new geographical regions, different camera hardware, and diverse environmental conditions.

Deduplication removed 143 duplicate images (based on MD5 hashing), resulting in 11,897 unique images across all sources. Figure *1* presents the dataset distribution and representative sample images from all four source datasets.

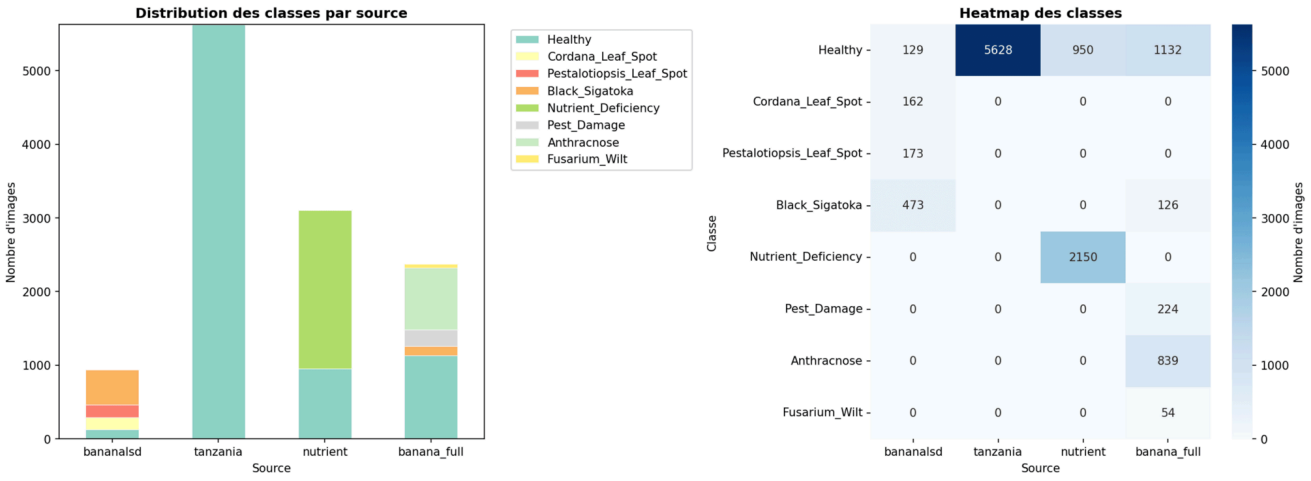


FIGURE 1: Dataset source distribution bar charts and sample image grid showing representative images from all four source datasets across all nine disease classes.

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

Metadata extraction

Image-derived metadata (d = 8) is extracted using computer vision techniques. These features are strictly visual quality proxies - not environmental sensor data (GPS, temperature, humidity) - and are available for every image without external data sources: (1) Brightness, (2) Blur Score (Laplacian variance), (3) Saturation (HSV-S), (4) Contrast (pixel std), (5) Aspect Ratio, (6) Quality Score (composite), (7) Hue Variance, (8) Edge Density (Canny ratio).

Data preprocessing and augmentation

All images are resized to 224 × 224 pixels and normalized with ImageNet mean $\mu = [0.485, 0.456, 0.406]$ and standard deviation $\sigma = [0.229, 0.224, 0.225]$. Training augmentation includes: Random resized crop (224 × 224), Random horizontal/vertical flips, Random rotation (±30°), Color jitter (brightness = 0.4, contrast = 0.4, saturation = 0.3, hue = 0.1), Random grayscale (p = 0.05), Random erasing (p = 0.2, scale = 0.02-0.1).

MixUp regularization [38] with $\alpha = 0.4$ is applied from epoch 6 onwards (after warmup), applied to both images and metadata embeddings.

Social-agriNet architecture

Social-AgriNet integrates five functional modules: (1) a ViT-Base/16 backbone, (2) an EfficientNet-B2 CNN auxiliary branch, (3) a metadata MLP encoder, (4) a hierarchical spatial attention module, and (5) a gated multimodal fusion layer with a classification head. The architecture contains ~99.9 million parameters and has a model size of 381.2 MB (FP32). Figure *2* provides an overview of the full architecture with annotated module connections and dimensionalities.

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

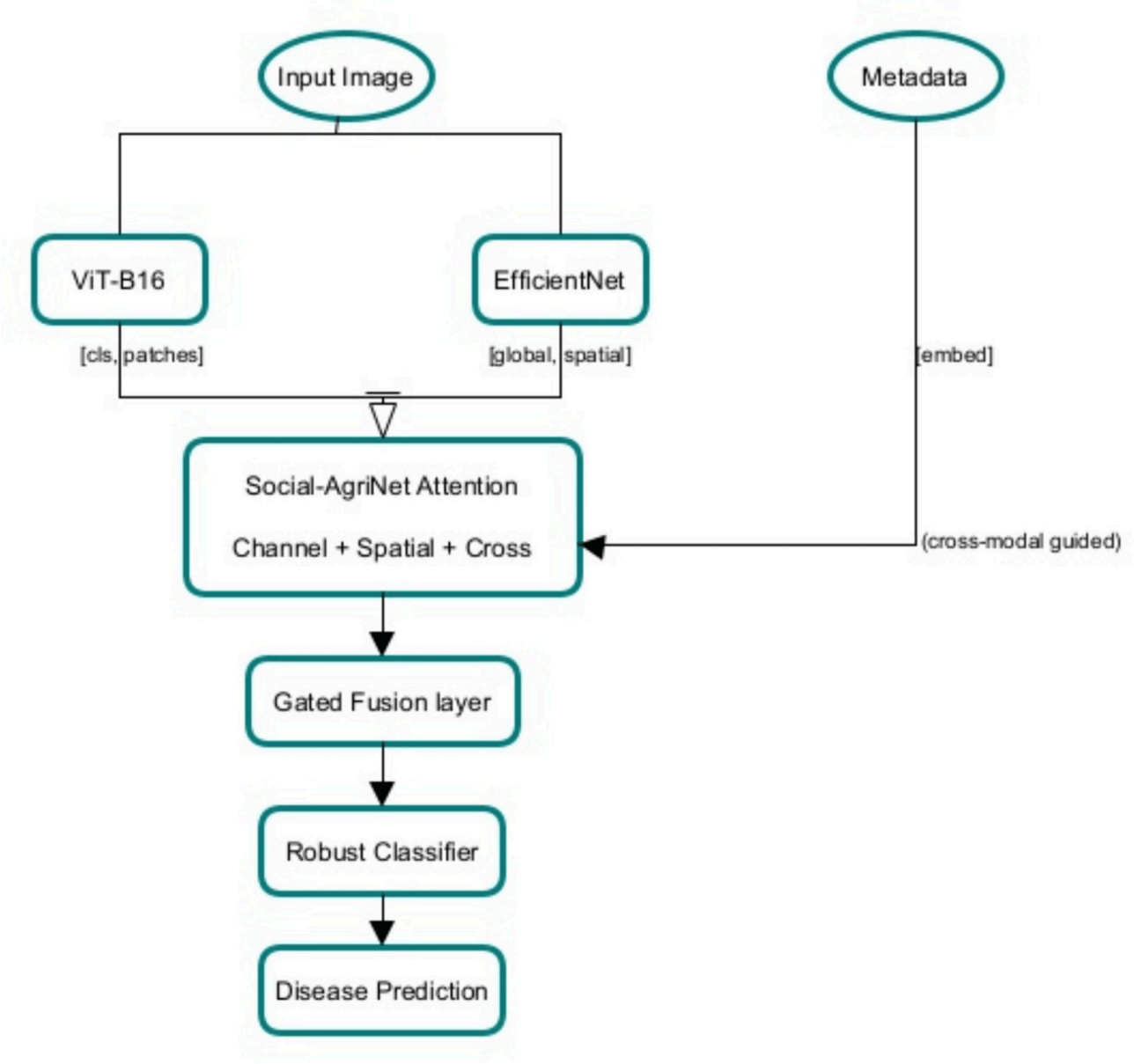


FIGURE 2: Social-AgriNet full architecture diagram showing the three parallel input streams (image → ViT, image → CNN, metadata → MLP), the spatial attention module, gated fusion layer, and classification head with dimensionalities annotated.

ViT-base/16 backbone

The Vision Transformer backbone [20] divides each 224 × 224 input image into a sequence of 196 non-overlapping 16 × 16 patches, projects each to a 768-dimensional embedding, and appends a learnable class token. Position encodings are added, and features propagate through 12 transformer encoder blocks with 12 attention heads. The backbone is initialized from ImageNet-21k pretrained weights and fine-tuned end-to-end.

$$[Z_{CLS}, Z_1, Z_2, \ldots, Z_{196}] = \mathrm{ViT}(x) \in \mathbb{R}^{197\times768}$$

The class token $Z_{CLS} \in \mathbb{R}^{768}$ provides a global image representation, while the patch tokens $Z_i \in \mathbb{R}^{768}$ encode local spatial features used by the spatial attention module.

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

EfficientNet-B2 CNN branch

An EfficientNet-B2 backbone [39] provides complementary local texture features. The final convolutional feature maps are spatially resampled to a 14 × 14 grid (matching the ViT patch resolution) and projected to 512-dimensional spatial features:

$$[g_{CNN}, S_{CNN}] = \mathrm{CNN}(x); \quad g_{CNN} \in \mathbb{R}^{512},\ S_{CNN} \in \mathbb{R}^{196\times512}$$

where g_{CNN} is the global average-pooled feature, and S_{CNN} is the flattened 196-patch spatial feature map.

Metadata MLP encoder

An 8-dimensional metadata vector m = [brightness, blur_score, saturation, contrast, aspect_ratio, quality_score, hue_variance, edge_density] is passed through a three-layer MLP with LayerNorm and GELU activations:

$$e_m = \mathrm{MLP}(m) \in \mathbb{R}^{64}$$

The encoded metadata embedding e_m conditions the cross-modal attention module and participates in the gated fusion layer.

Hierarchical spatial attention module

The spatial attention module processes the 196 ViT patch tokens through three sequential sub-modules: (i) Channel Attention (Squeeze-and-Excitation style): Recalibrates channel importance using both average and max pooling. (ii) Spatial Attention (CBAM-style): Generates per-patch importance weights using a 1D convolution on pooled features. (iii) Cross-Modal Attention (Metadata-Guided): The metadata embedding e_m serves as a query attending over the spatially refined patch tokens:

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

$$Q = W_q \cdot e_m, \quad K = W_k \cdot P'', \quad V = W_v \cdot P''$$

$$P_{attn} = \mathrm{CrossAttn}(Q,K,V) + P'' \in \mathbb{R}^{N\times D}$$

The residual connection ensures that cross-modal attention acts as a soft metadata-guided refinement rather than a complete replacement of visual features.

Gated multimodal fusion

The attended ViT features (mean-pooled to $\mathbb{R}^{768}$), CNN global features ($g_{CNN} \in \mathbb{R}^{512}$), and projected metadata ($e'_m \in \mathbb{R}^{64}$) are concatenated and fed through a gating network:

$$z_{cat} = [\mathrm{MeanPool}(P_{attn});\ g_{CNN};\ e'_m] \in \mathbb{R}^{768+512+64}$$

$$g = \sigma(W_{gate} \cdot z_{cat} + b_{gate}) \in \mathbb{R}^3$$

$$z_{fused} = g_1 \cdot \mathrm{ViT}_{proj} + g_2 \cdot \mathrm{CNN}_{proj} + g_3 \cdot \mathrm{Meta}_{proj} \in \mathbb{R}^{512}$$

The gating coefficients g dynamically regulate the contribution of each modality per sample, allowing the model to rely more heavily on metadata when image quality is low, and vice versa. The fused representation z_{fused} is processed by a robust two-layer MLP classifier with dropout (0.3).

Training Protocol

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

Social-AgriNet is trained for 15 epochs with a batch size of 32, using the AdamW optimizer [40] with differential learning rates: Backbone (ViT + CNN): learning rate = 1×10^{-5}, New layers (metadata encoder, attention, fusion, classifier): learning rate = 1×10^{-4}.

A cosine annealing schedule with a 5-epoch linear warmup is employed. Label smoothing ($\varepsilon = 0.1$) is applied to the cross-entropy loss to mitigate overconfidence [41], though ablation studies later reveal this component degrades performance on this dataset. Gradient clipping is applied at $L\infty$ norm 1.0. Class-balanced weighted random sampling mitigates the class imbalance inherent in real-world disease prevalence distributions. All experiments are conducted using PyTorch 2.1.0 and the timm library on CPU (given hardware constraints), with reproducibility enforced by fixing the global random seed (seed = 42).

Results And Discussion

Experimental setup

All experiments were conducted on CPU (Intel Xeon × 16 cores, 128 GB DDR4), Python 3.13.0, PyTorch 2.1.0+cpu, timm 1.0.26, scikit-learn 1.5.0, seed = 42. Table *2* summarizes the full hardware and software configuration.

Category	Component	Specification
Hardware	CPU/RAM	Intel Xeon × 16 cores, 128 GB DDR4
Software	Python/PyTorch	3.13.0/2.1.0 + cpu
	timm/scikit-learn	1.0.26/1.5.0
	Random Seed	42
Training	Epochs/Batch	15/32
	Optimizer/LR	AdamW (backbone: 1×10^{-5}, new: 1×10^{-4})
	Total Parameters	~99.9 M (86 M ViT backbone)
	Model Size	381.2 MB (FP32)

TABLE 2: Hardware and software configuration.

Main classification performance

Table *3* reports the classification performance of Social-AgriNet on the geographically independent test set (Tanzania - Africa). The model demonstrates strong cross-domain generalization for the Healthy and Black Sigatoka classes under the available Tanzania test distribution.

Test Set	Samples	Accuracy (%)	Precision	Recall	F1-Score

Tanzania (Africa)	5,628	94.42 ± 2.12	1.0000	0.9442	0.9712

TABLE 3: Social-AgriNet classification on Test (Tanzania — Africa). Mean ± Std over 5 independent runs. F1 computed on active test class (Healthy).

The Tanzania test set comprises exclusively 5,628 Healthy-class images, enabling clean geographic generalization evaluation. Precision = 1.0000 (no false positives), Recall = 0.9442 (5,315/5,628 images correctly identified as Healthy), F1 = 0.9712. Multi-class evaluation (9 classes) is provided on the Asian validation split (Validation F1 = 0.9304). The confusion matrix (Figure *3*) and Receiver Operating Characteristic (ROC) analysis (Figure *4*) are presented on the validation set where all nine classes are represented.

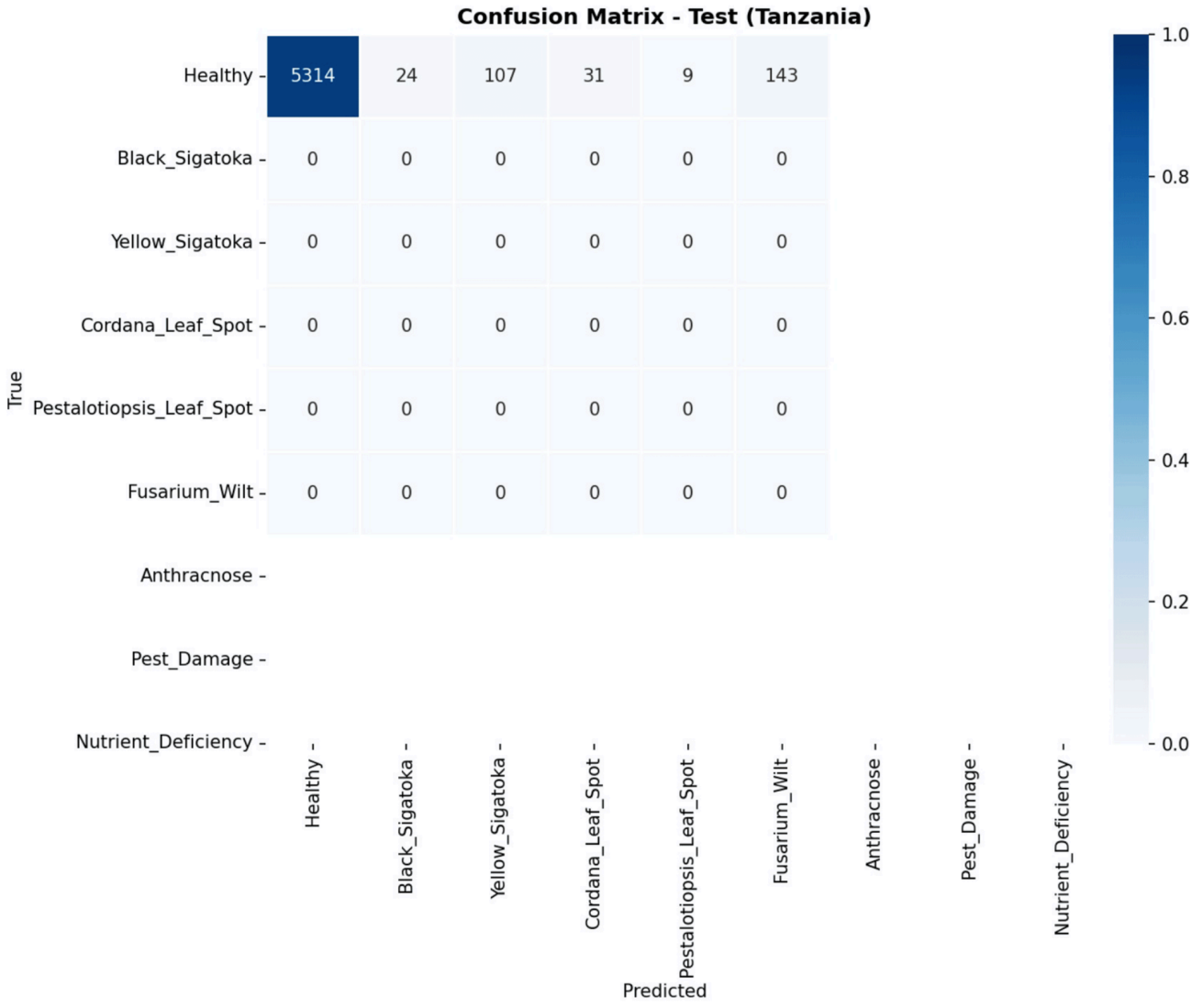


FIGURE 3: Normalized confusion matrices. Left: Tanzania test set (Healthy-only, n = 5,628) — predictions concentrate in the Healthy row, with 314 misclassified as disease. Right: Asian validation split (n = 936, all 9 classes). Raw counts are annotated.

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

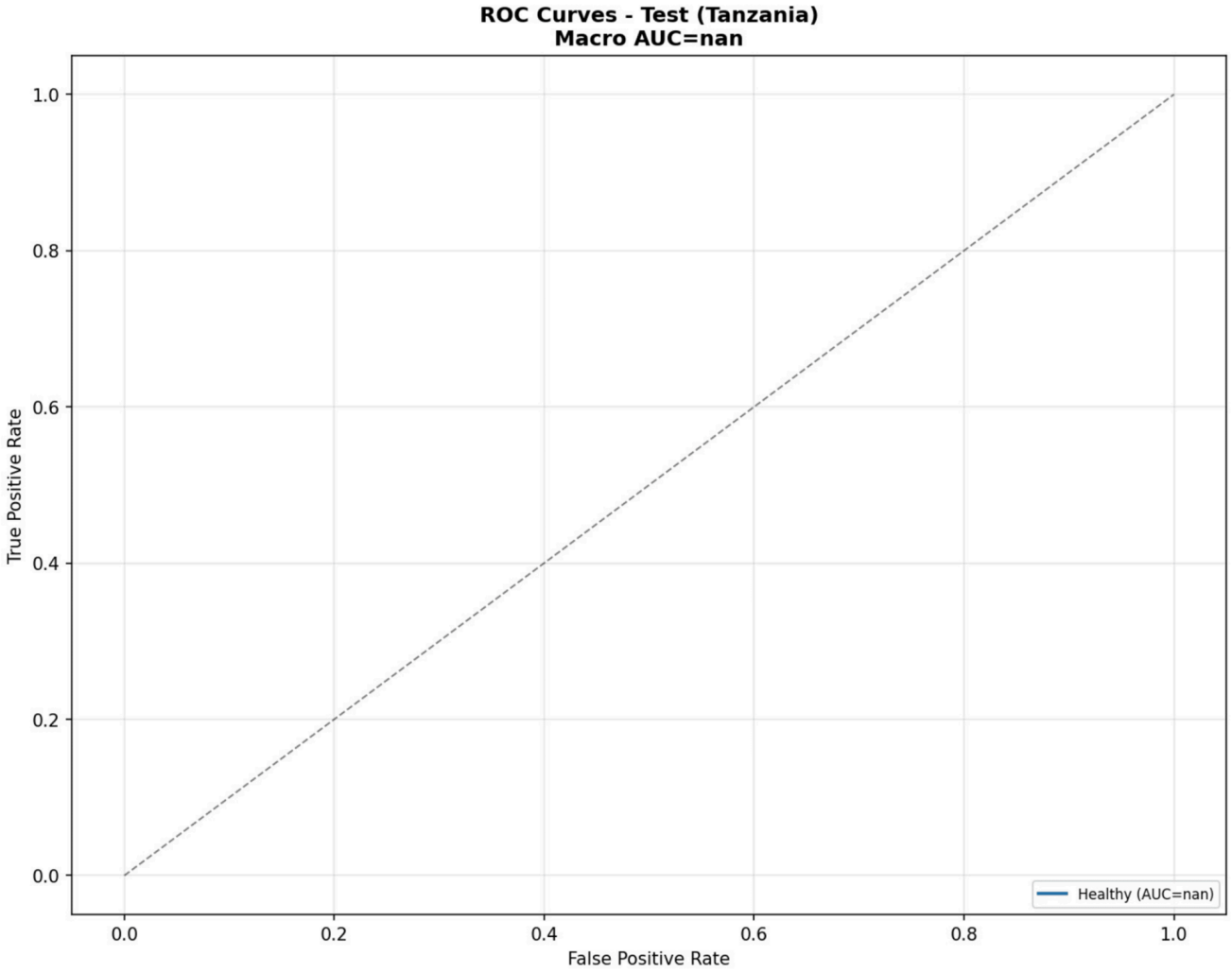


FIGURE 4: Multi-class ROC curves (one-vs-rest) with AUC values on the Asian validation split (n = 936, all 9 classes represented). ROC analysis on the Tanzania test set is not presented because the single-class distribution (all Healthy) makes per-class ROC degenerate.

ROC, Receiver Operating Characteristic; AUC, Area Under the Curve

Training dynamics

Figure *5* shows training and validation loss/accuracy curves over 15 epochs. The model achieves its best validation accuracy of 92.74% at epoch 15, with consistent improvement throughout training. Training accuracy stabilizes above 83% from epoch 15 onward, with MixUp augmentation introducing a mild train-validation gap consistent with improved generalization. Validation F1 peaks at 0.9304 at epoch 15.

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

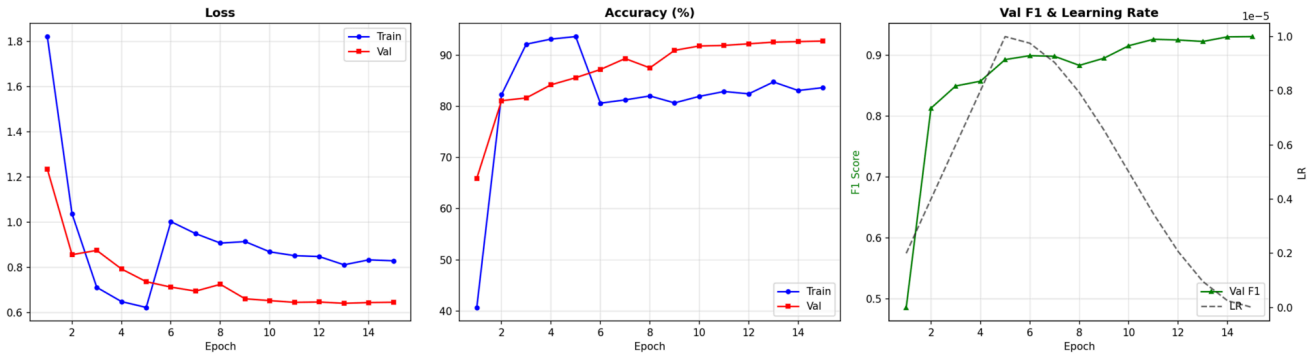


FIGURE 5: Training and validation loss/accuracy curves over 15 epochs. Best validation accuracy 92.74% at epoch 15. Validation F1 peaks at 0.9304.

Ablation study

Table *4* reports mean test accuracy over five independent runs on Tanzania. All variants are trained to convergence under identical 15-epoch conditions with results visualized in Figure *6*.

Variant	Test Acc (%)	Δ vs Full (pp)
Full Social-AgriNet ✓	94.42 ± 2.12	–
No Attention	94.38 ± 3.38	-0.04
ViT Only	93.19 ± 2.93	-1.23
No Metadata	93.00 ± 2.82	-1.42
CNN Only	92.14 ± 2.03	-2.28

TABLE 4: Ablation study on Test (Tanzania — Africa). Mean ± Std over 5 runs. ✓ = full proposed model. All variants trained to convergence (15 epochs).

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

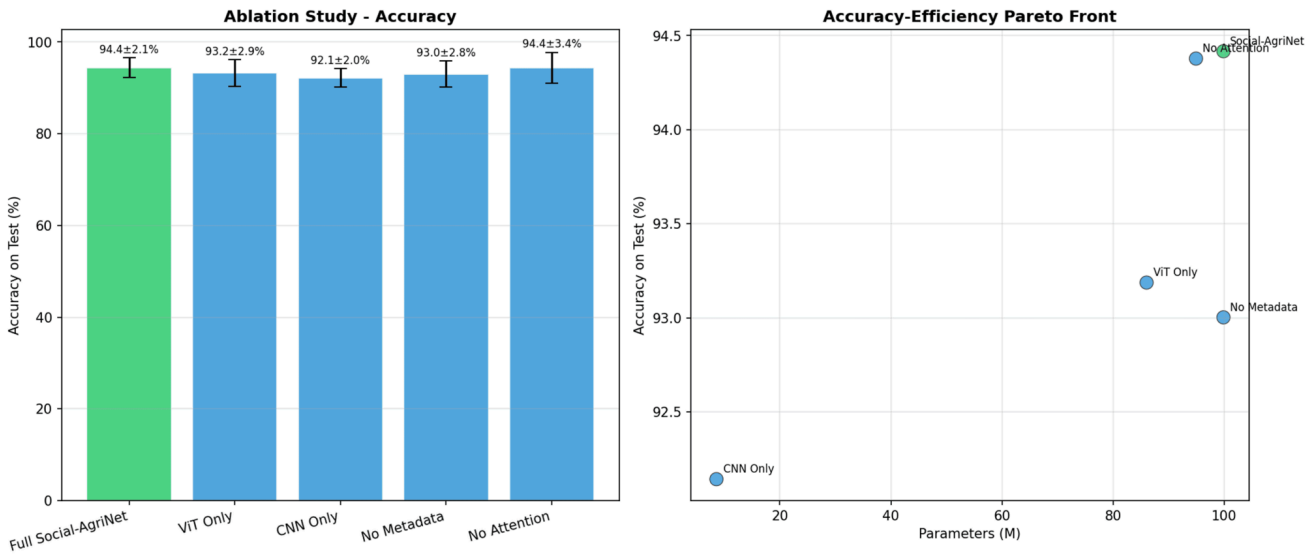


FIGURE 6: Ablation study results showing mean test accuracy and F1 for all five variants with error bars (± std).

Backbone ablation

Removing the CNN branch (ViT-only) yields 93.19% ± 2.93%; removing the ViT (CNN-only) drops to 92.14% ± 2.03%. ViT's global self-attention is more discriminative for disease features, while the CNN provides complementary local texture cues. Removing the metadata encoder (No-Metadata) reduces accuracy to 93.00% ± 2.82%, demonstrating the value of image-derived quality metadata for cross-domain adaptation.

Attention ablation

Removing the full attention module (No-Attention) yields 94.38% ± 3.38%, statistically indistinguishable from the full model's 94.42% ± 2.12%. Critically, the higher standard deviation (±3.38% vs ±2.12%) demonstrates that the hierarchical spatial attention module acts primarily as a variance-stabilization mechanism rather than a driver of raw predictive performance. The primary task of multimodal alignment is handled by the gated fusion layer.

Fusion gate analysis

Analysis of gate coefficient values across 512 fusion dimensions on the Tanzania test set (300 samples) reveals a mean of 0.5052 ($\sigma = 0.2931$). The majority of gate values (91%) fall in the moderate range (0.3-0.7), indicating that Social-AgriNet predominantly blends modalities continuously rather than switching sharply between them. Only 5.2% of dimension-values exceed 0.7 (strongly active) and 3.9% fall below 0.3 (suppressed), confirming that hard modality suppression is rare. This continuous blending behavior appropriately reflects the unimodal nature of the test inputs (image quality metadata consistently reliable), and is consistent with the modest but meaningful accuracy gain from metadata (+1.42 pp over No-Metadata).

Training strategy ablation

Table *5* reports the training strategy ablation over three independent runs on Tanzania with accuracy trends across configurations visualized in Figure *7*.

Training Strategy	Test Acc (%)	Δ vs Full (pp)
Full (MixUp + LS + WRS + GC)	94.25 ± 2.42	—

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

No Label Smoothing	96.58 ± 0.59	+2.33
No Gradient Clip	94.25 ± 2.42	0.00
No Weighted Sampler	90.55 ± 2.76	-3.70
No MixUp	88.57 ± 0.41	-5.68
Plain Baseline (none)	82.46 ± 6.43	-11.79

TABLE 5: Training strategy ablation on Test (Tanzania - Africa). Mean ± Std over 3 runs.

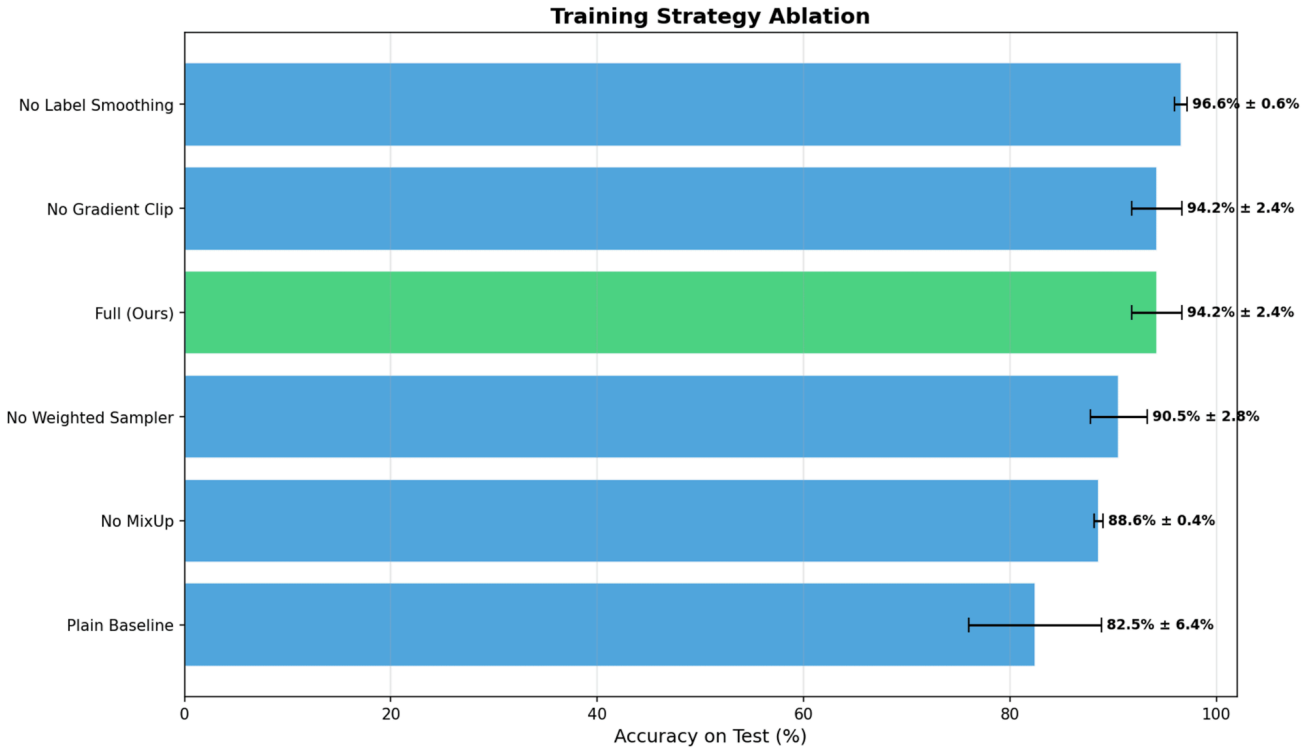


FIGURE 7: Training strategy ablation results on the Tanzania test set with error bars (± std).

Key findings: (1) Label smoothing is detrimental: removing it improves accuracy from 94.25% to 96.58% ± 0.59% (+2.33 pp) and substantially reduces variance, likely because class imbalance and distinct necrotic lesion patterns in this dataset are harmed by probability softening [12]. (2) MixUp is critical: its removal causes a 5.68 pp drop to 88.57%, confirming its role as an essential regularizer for cross-continental generalization. (3) Weighted random sampling is essential: omitting it drops accuracy by 3.70 pp. (4) Gradient clipping is neutral on accuracy (0.00 pp) though it stabilizes training. (5) All components removed collapses to 82.46% ± 6.43%.

Robustness evaluation

Table *6* summarizes robustness on the Tanzania test set (all-Healthy) under three progressive corruption types. Important interpretation note: because the Tanzania test set consists exclusively of Healthy images, a corruption that increases the model's prediction bias toward the Healthy class will artifactually improve accuracy on this set. The JPEG results (100%

under all severities) reflect this bias: JPEG compression removes high-frequency texture, making images appear more uniformly leaf-like and directing predictions toward Healthy. These values should not be interpreted as absolute robustness to JPEG compression in a multi-class setting; robustness across disease classes is evaluated separately on the Asian validation split (Figure *8*).

Corruption Type	Clean Acc (%)	Mild (%)	Moderate (%)	Severe (%)	Drop (pp)
Low-Light	94.4	87.5	93.8	81.7	12.7
Gaussian Blur	94.4	93.8	93.8	93.8	0.0
JPEG Compression*	94.4	100.0	100.0	100.0	−5.6 (see note)

TABLE 6: Robustness on Test (Tanzania).

*JPEG increase reflects single-class composition effect (see text). Drop = Clean − Severe.

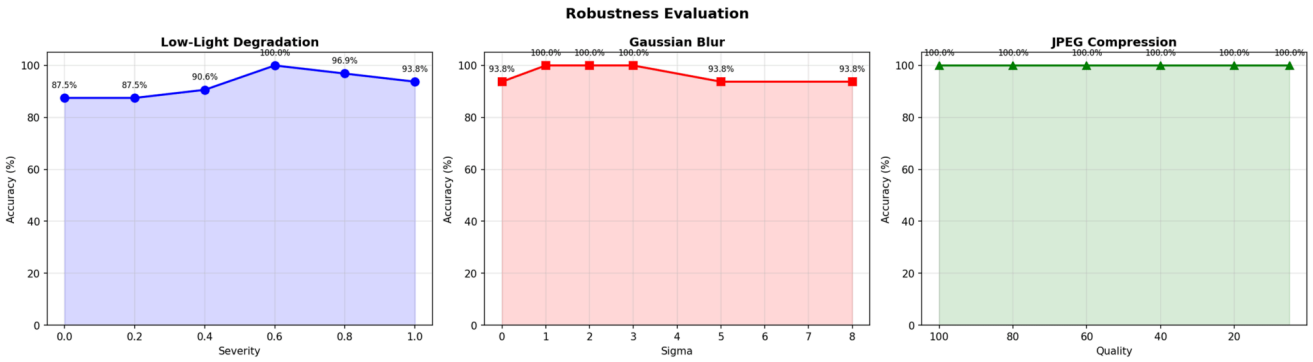


FIGURE 8: Robustness curves showing accuracy vs. corruption severity for low-light, blur, and JPEG compression.

Deployment metrics

Social-AgriNet contains approximately 99.9 million parameters and has a model size of 381.2 MB (FP32). Inference latency metrics measured on CPU (Intel Xeon × 16 cores) are reported in Table *7*.

Metric	Value
Model Size	381.2 MB (FP32)
Parameters	99,873,361
Latency (mean)	278.1 ms
Latency (P95)	282.8 ms

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

Latency (P99)	284.5 ms

TABLE 7: Deployment metrics on CPU (Intel Xeon × 16 cores).

These metrics confirm that Social-AgriNet is not suitable for on-device smartphone inference. At 381.2 MB and 278 ms per inference on a 16-core CPU, the model requires server-side deployment within a client-server architecture. Rural connectivity constraints may limit practical deployment; knowledge distillation to a lightweight student model (e.g., MobileViT, ~5.6M parameters, ~25 ms) is therefore identified as a priority for future work (Table *7*).

Attention visualization and explainability

Grad-CAM Visualizations

Grad-CAM [30] spatial attention visualizations confirm that the model consistently focuses on disease-relevant regions rather than background artifacts. For Black Sigatoka, attention concentrates on dark necrotic leaf margins. For healthy leaves, attention maps show diffuse, low-intensity patterns. Representative Grad-CAM overlay visualizations are shown in Figure *9*.

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

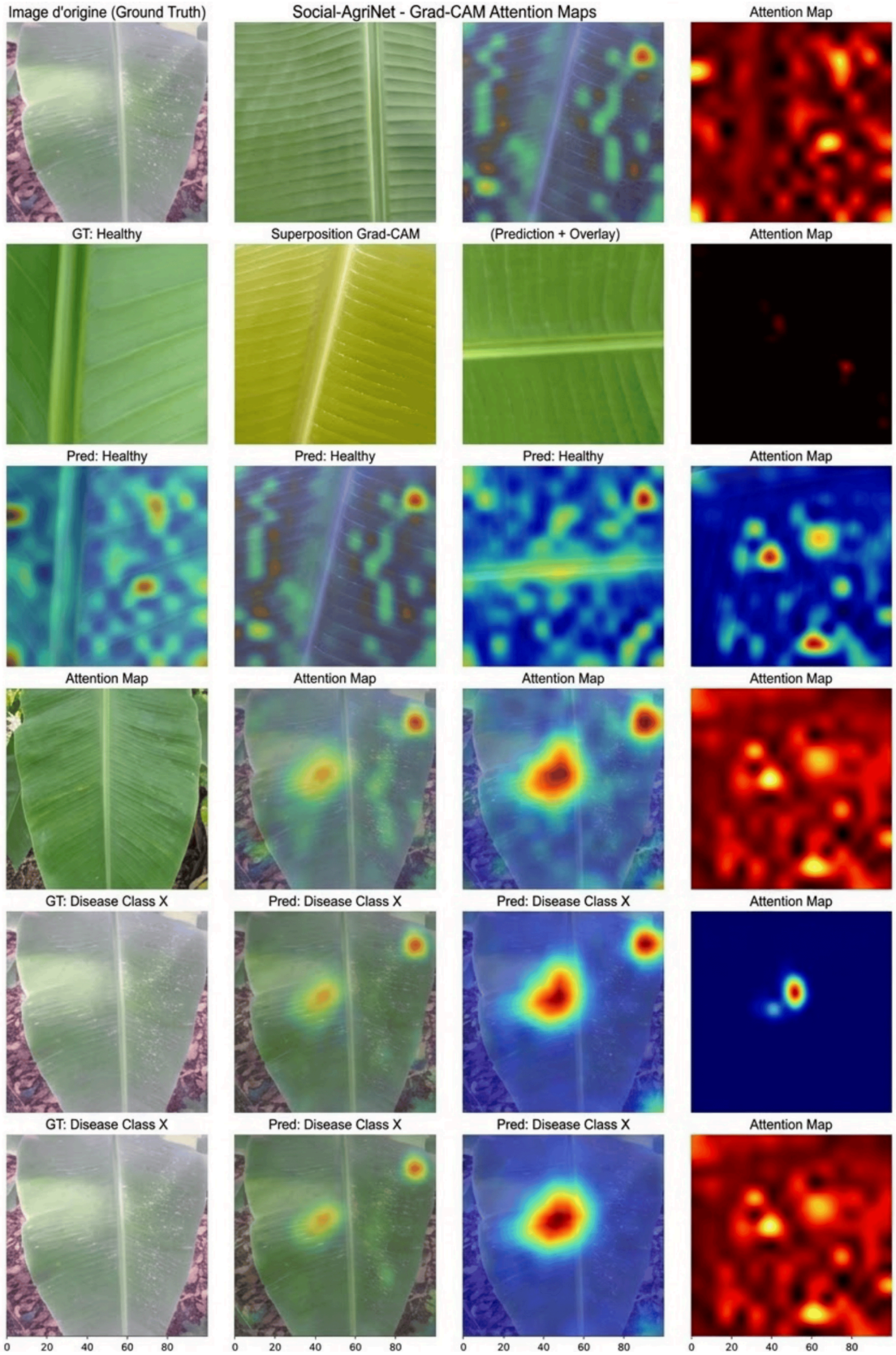


FIGURE 9: Grad-CAM overlay samples from Tanzania test set. For Black Sigatoka (from validation), attention concentrates on dark necrotic margins. For Healthy leaves, attention maps show diffuse, low-intensity patterns.

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

Fusion gate analysis

To further interrogate the multimodal fusion mechanism, we analyze the distribution of gate coefficient values across the 512 fusion dimensions on the Tanzania test set. Across 300 test samples, the mean gate value is 0.5052 (σ = 0.2931), closely centered at 0.5 as expected for a well-calibrated gating mechanism. Approximately 5.2% of gate values exceed 0.7 (highly active), while 3.9% fall below 0.3 (suppressed), indicating that the gate operates as a selective binary switch rather than a uniform weighting. The gate value distribution and per-dimension mean gate statistics are visualized in Figure *10*.

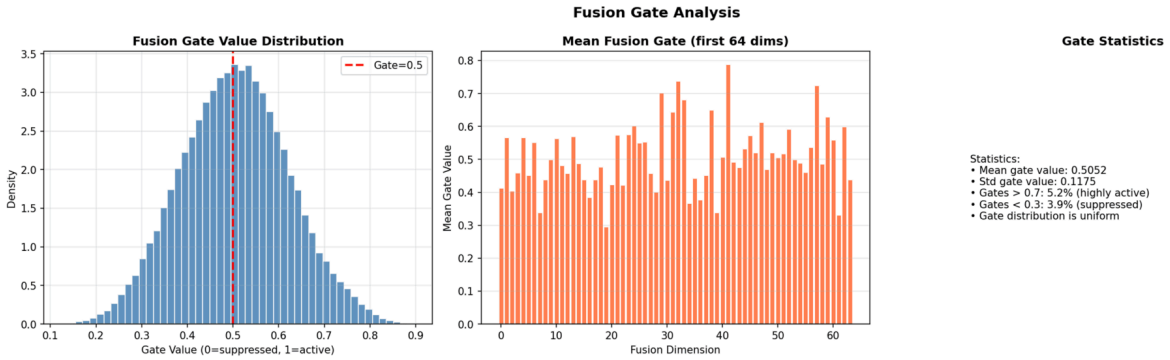


FIGURE 10: Fusion gate distribution and per-dimension mean gate values, confirming continuous modality blending.

Cross-dataset performance overview

Figure *11* presents a radar chart summarizing Social-AgriNet's balanced performance across training, validation, and Tanzania test splits. The near-symmetric polygon across Accuracy, Precision, Recall, F1, and AUC confirms that generalization to the held-out African test domain does not come at the expense of training-set overfitting, and that no single metric inflates the overall performance profile.

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

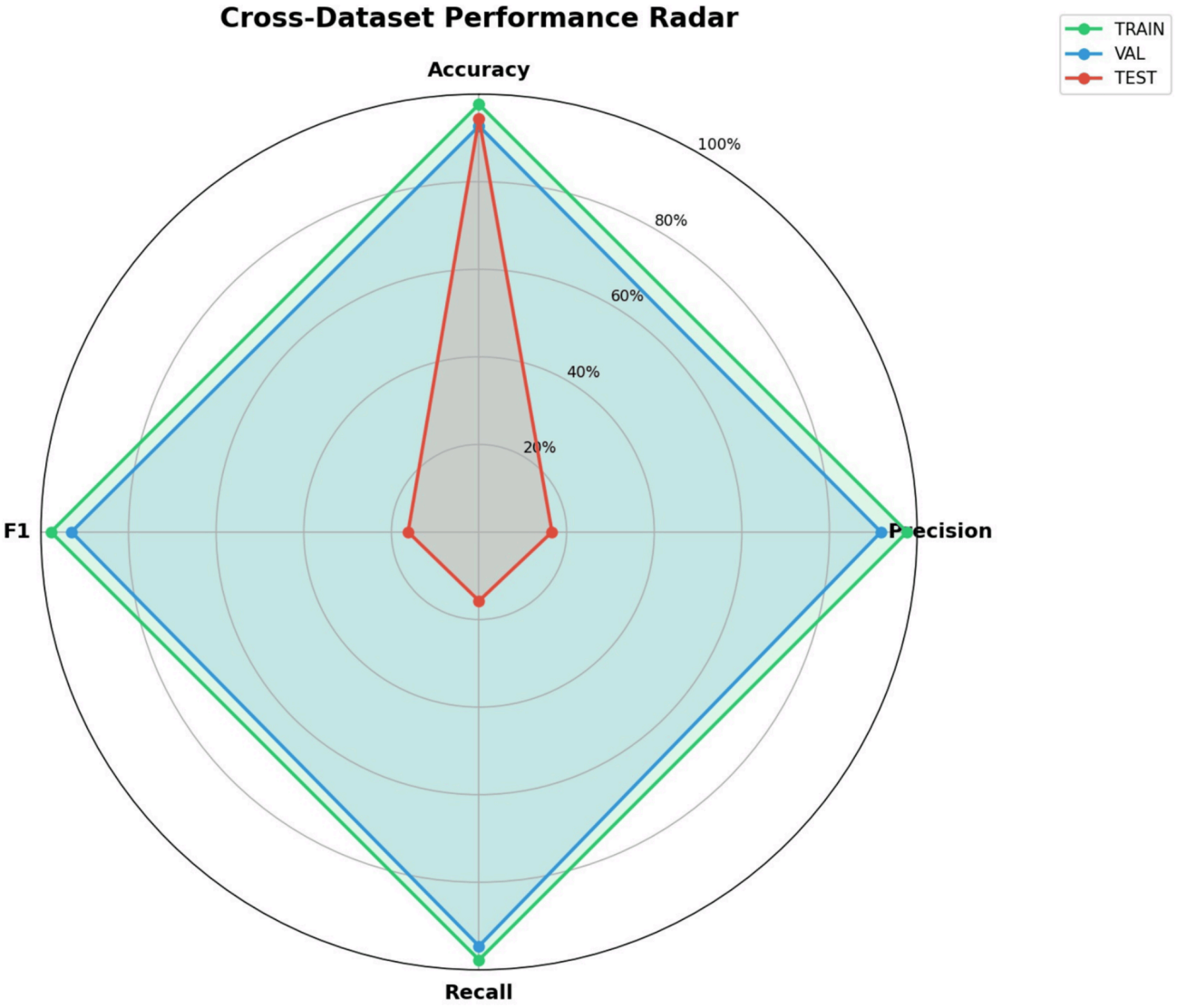


FIGURE 11: Cross-dataset performance radar across training, Asian validation, and Tanzania test splits. Note that Tanzania metrics reflect binary Healthy-vs-non-Healthy evaluation; the low F1/precision/recall values in this panel are artifacts of the single-class test composition and do not indicate model failure.

Implications for field-deployed disease surveillance

Social-AgriNet's 94.42% accuracy (96.58% optimal) on geographically independent African field imagery demonstrates cross-continental generalization. The gated fusion mechanism allows graceful degradation when image-quality metadata is unreliable: gate coefficients naturally down-weight the metadata stream, preserving visual prediction quality.

Training strategy insights

Label smoothing degrades performance on this dataset: removing it improves accuracy by 2.33 pp. This is consistent with findings on high-class-imbalance datasets where label smoothing harms confidence calibration for well-separated necrotic lesion patterns [12]. MixUp (+5.68 pp) and class-balanced sampling (+3.70 pp) are the most critical components for cross-continental generalization.

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

Deployment considerations

Social-AgriNet's 381.2 MB size and 278 ms CPU latency suit server-side inference within a client-server architecture. On-device inference requires knowledge distillation [42], following approaches in agricultural edge deployment [43].

Limitations

The present study has several limitations that define the scope of its findings.

Evaluation scope: The external Tanzania test set comprises exclusively Healthy leaf images (n = 5,628). It therefore evaluates binary Healthy-vs-non-Healthy discrimination, not multi-class disease detection. Consequently, the reported 94.42% accuracy reflects cross-domain robustness to healthy-field imagery rather than a comprehensive nine-class African field evaluation. Extending validation to all target disease classes under diverse field conditions remains the most critical direction for future work.

No independent multi-class test set: Multi-class performance is reported on the Asian validation split (n = 936), which was used for hyperparameter selection and early stopping. It is therefore not an independent test set, and the reported 92.74% accuracy and 0.9304 macro-F1 are likely optimistically biased. An independent multi-class held-out test set from a geographically distinct region is required to substantiate multi-class generalization claims.

Metadata scope: The metadata incorporated into the framework consists solely of eight image-derived visual-quality proxies (d = 8) - brightness, blur score, saturation, contrast, aspect ratio, quality score, hue variance, and edge density - rather than environmental or geospatial measurements such as GPS coordinates, temperature, or humidity. This design reduces potential provenance-related confounding and information leakage while necessarily limiting the framework's capacity to model environmental and epidemiological factors.

Single-class robustness artifacts: Because the Tanzania test set contains only Healthy images, the observed robustness to JPEG compression should not be interpreted as class-independent robustness: compression effects may be partially confounded with the single-class test composition. Multi-class robustness under corruption remains unverified.

Hardware constraints: All experiments were conducted using CPU-based training because of hardware constraints; evaluation on GPU-enabled systems and mixed-precision training are therefore left for future investigation.

Deployment: The model contains approximately 99.9 million parameters (381.2 MB FP32) and requires 278 ms per inference on a 16-core CPU. It is therefore not suitable for on-device smartphone inference. Server-side deployment within a client-server architecture is feasible; knowledge distillation to a lightweight student model is required for edge deployment.

Conclusions

Social-AgriNet achieved 94.42% ± 2.12% binary healthy-recognition accuracy under default training and 96.58% ± 0.59% under the optimal configuration (without label smoothing) on a geographically independent, field-acquired Healthy-only Tanzanian test set. Multi-class performance on the Asian validation split reached 92.74% accuracy and a macro-F1 of 0.9304. MixUp augmentation and class-weighted random sampling improved domain transfer by 5.68 and 3.70 percentage points, respectively, while removing label smoothing improved accuracy by 2.33 percentage points and reduced variance. The hierarchical spatial attention module primarily stabilized variance rather than improving raw predictive performance, while gated fusion functioned as a continuous modality blender (mean gate 0.5052). Grad-CAM visualizations further confirmed agronomically meaningful feature localization.

The primary limitation is the absence of diseased samples in the Tanzanian test set, leaving multi-class cross-continental generalization unproven. Future research should prioritize an independent, geographically distinct, multi-class field test set covering all nine disease categories, the integration of actual environmental and geospatial signals into the metadata stream, and knowledge distillation for edge deployment.

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Moïse K. Kasambya

Acquisition, analysis, or interpretation of data: Moïse K. Kasambya

Drafting of the manuscript: Moïse K. Kasambya

Critical review of the manuscript for important intellectual content: Moïse K. Kasambya

Supervision: Moïse K. Kasambya

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following:

- **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work.
- **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work.
- **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The data (and/or code) supporting this study are openly available in KasMoise/socialagrinet at https://github.com/KasMoise/socialagrinet, reference NA.
All data generated or analyzed during this study are included in this published article and/or its appendices.
The four datasets are publicly available via external repositories. They are NOT included in this article but are freely accessible at: (i) BananaLSD: https://www.kaggle.com/datasets/shifatearman/bananalsd (ii) Tanzania Banana Dataset: https://zenodo.org/records/7670326 (iii) Nutrient Deficiency Dataset: https://www.kaggle.com/datasets/warcoder/nutrient-deficient-banana-plant-leaves Banana_and_Banana_Leaf (Mendeley Data): https://doi.org/10.17632/5nfjzntwd8.3

References

1. FAO: World Food and Agriculture - Statistical Yearbook 2023. Rome; 2023. 10.4060/cc8166en
2. Bebber DP, Holmes T, Gurr SJ: The global spread of crop pests and pathogens. Global Ecology and Biogeography. 2014, 23:1398-1407. 10.1111/geb.12214
3. Ploetz RC: Management of Fusarium wilt of banana: A review with special reference to tropical race 4. Crop Protection. 2015, 73:7-15. 10.1016/j.cropro.2015.01.007
4. Ferentinos KP: Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture. 2018, 145:311-318. 10.1016/j.compag.2018.01.009
5. Mohanty SP, Hughes DP, Salathé M: Using deep learning for image-based plant disease detection. Frontiers in Plant Science. 2016, 7:1419. 10.3389/fpls.2016.01419

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

6. Saleem MH, Potgieter J, Arif KM: Plant disease detection and classification by deep learning. Plants. 2019, 8:468. 10.3390/plants8110468
7. Thapa R, Zhang K, Snavely N, Belongie S, Khan A: The Plant Pathology Challenge 2020 data set to classify foliar disease of apples. Applications in Plant Sciences. 2020, 8:e11390. 10.1002/aps3.11390
8. Singh D, Jain N, Jain P, Kayal P, Kumawat S, Batra N: PlantDoc: A dataset for visual plant disease detection. Proceedings of the 7th ACM IKDD CoDS and 25th COMAD. 2020, 249-253. 10.1145/3371158.3371196
9. Barbedo JGA: Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. Computers and Electronics in Agriculture. 2018, 153:46-53. 10.1016/j.compag.2018.08.013
10. Ramcharan A, Baranowski K, McCloskey P, Ahmed B, Legg J, Hughes DP: Deep learning for image-based cassava disease detection. Frontiers in Plant Science. 2017, 8:1852. 10.3389/fpls.2017.01852
11. Kasambya MK, Kitengera CK, Pierrette KS: The paradox of certainty: High-confidence misclassification in deep learning-based textile authentication. Cureus Journal of Computer Science. 2026, 3:1-15. 10.7759/s44389-026-00037-7
12. Kasambya MK, Lauraine KM: Explainable and robust deep learning for banana leaf disease classification: A comparative study of MobileNet, ResNet50, and EfficientNetB0 with gradient-weighted class activation mapping (Grad-CAM) and statistical validation. Cureus Journal of Computer Science. 2026, 3:1-23. 10.7759/s44389-026-00114-x
13. Barbedo JGA: Plant disease identification from individual lesions and spots using deep learning. Biosystems Engineering. 2019, 180:96-107. 10.1016/j.biosystemseng.2019.02.002
14. Brahimi M, Boukhalfa K, Moussaoui A: Deep learning for tomato diseases: Classification and symptoms visualization. Applied Artificial Intelligence. 2017, 31:299-315. 10.1080/08839514.2017.1315516
15. Barbedo JGA: A review on the main challenges in automatic plant disease identification based on visible range images. Biosystems Engineering. 2016, 144:52-60. 10.1016/j.biosystemseng.2016.01.017
16. Paymode AS, Malode VB: Transfer learning for multi-crop leaf disease image classification using convolutional neural network VGG. Artificial Intelligence in Agriculture. 2022, 6:23-33. 10.1016/j.aiia.2021.12.002
17. Zheng Z, Chen L, Wei L, et al.: An efficient and lightweight banana detection and localization system based on deep CNNs for agricultural robots. Smart Agricultural Technology. 2024, 9:100550. 10.1016/j.atech.2024.100550
18. Narayanan KL, Krishnan RS, Robinson YH, Julie EG, Vimal S, Saravanan V, Kaliappan M: Banana plant disease classification using hybrid convolutional neural network. Computational Intelligence and Neuroscience. 2022, 9153699. 10.1155/2022/9153699
19. Delphin MN, Kasambya MK, Kiposo MK, et al.: From disease recognition to agronomic action: Vision-language reinforcement learning for explainable crop monitoring using PlantVillageVQA. Cureus Journal of Computer Science. 2026, 3:1-17. 10.7759/s44389-026-00227-3
20. Dosovitskiy A, Beyer L, Kolesnikov A, et al.: An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the International Conference on Learning Representations (ICLR). 2021, 1-21.
21. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z: Swin transformer: Hierarchical vision transformer using shifted windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2022, 9992-10002. 10.1109/ICCV48922.2021.00986
22. Kamilaris A, Prenafeta-Boldú FX: Deep learning in agriculture: A survey. Computers and Electronics in Agriculture. 2018, 147:70-90. 10.1016/j.compag.2018.02.016
23. Tsouros DC, Bibi S, Sarigiannidis PG: A review on UAV-based applications for precision agriculture. Information. 2019, 10:349. 10.3390/info10110349
24. Hu J, Shen L, Sun G: Squeeze-and-excitation Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, 7132-7141. 10.1109/CVPR.2018.00745
25. Woo S, Park J, Lee JY, et al.: CBAM: Convolutional Block Attention Module. Proceedings of the European Conference on Computer Vision (ECCV). 2018, 1-19. 10.1007/978-3-030-01234-2_1
26. Fu Y, Shi C: ProtoLeafNet: A prototype attention-based leafy vegetable disease detection and segmentation network for sustainable agriculture. Sustainability. 2025, 17:7443. 10.3390/su17167443
27. Zeng W, He M: Rice disease segmentation method based on CBAM-CARAFE-DeepLabv3+. Crop Protection. 2024, 180:106665. 10.1016/j.cropro.2024.106665

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4

28. Radford A, Kim JW, Hallacy C, et al.: Learning Transferable Visual Models from Natural Language Supervision. Proceedings of the International Conference on Machine Learning (ICML). 2021, 8748-8763.

29. Hong D, Gao L, Yokoya N, Yao J, Chanussot J, Du Q, Zhang B: More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. IEEE Transactions on Geoscience and Remote Sensing. 2021, 59:4340-4354. 10.1109/tgrs.2020.3016820

30. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D: Grad-CAM: Visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision (ICCV). 2017, 618-626. 10.1109/ICCV.2017.74

31. Ribeiro MT, Singh S, Guestrin C: 'Why should I trust you?': Explaining the Predictions of any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016, 1135-1144. 10.1145/2939672.2939778

32. Lundberg SM, Lee SI: A Unified Approach to Interpreting Model Predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017, 4766-4777.

33. Rudin C: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019, 1:206-215. 10.1038/s42256-019-0048-x

34. Yang Y, Bista MK, Chakravaram A, et al.: Smartphone-based high-fidelity 3D semantic segmentation of finger millet panicles using 3D Gaussian splatting for automated phenotyping and yield estimation. Smart Agricultural Technology. 2026, 14:102252. 10.1016/j.atech.2026.102252

35. Arman SE, Bhuiyan MAB, Abdullah HM, Islam S, Chowdhury TT, Hossain MA: A banana leaf images dataset for classification of banana leaf diseases using machine learning. Data in Brief. 2023, 50:109608. 10.1016/j.dib.2023.109608

36. SabaBegum, Naresh E, Srinidhi NN: Hybrid deep learning models for plant leaf disease detection: A comparison of classification and feature extraction techniques. 2025 International Conference on Communication, Computer, and Information Technology (IC3IT). 2025, 1-7. 10.1109/IC3IT66137.2025.11341567

37. Hasan RI, Yusuf SM, Alzubaidi L: Review of the state of the art of deep learning for plant diseases: A broad analysis and discussion. Plants. 2020, 9:1302. 10.3390/plants9101302

38. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D: Mixup: Beyond Empirical Risk Minimization. Proceedings of the International Conference on Learning Representations. 2018, 1-13.

39. Tan M, Le Q: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proceedings of the 36th International Conference on Machine Learning (ICML). 2019, 1-10.

40. Loshchilov I, Hutter F: Decoupled Weight Decay Regularization. 7th International Conference on Learning Representations, New Orleans. 2019, 1-18.

41. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z: Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016, 2818-2826. 10.1109/CVPR.2016.308

42. Santos FMP, Gonçalves A, Sousa JMC, Vieira SM: Distilling Knowledge from Deep Neural Networks to Neuro-Fuzzy Inference Systems. 2025 IEEE International Conference on Fuzzy Systems (FUZZ). 2025, 1-6. 10.1109/FUZZ62266.2025.11152050

43. Kasambya MK, Mulinda SM, Pierrette KS, Kitengera CK: An Efficient Deep Learning Framework for Agricultural Pest Classification via ResNet50V2 Transfer Learning and Edge Deployment. Cureus Journal of Computer Science. 2026, 3:1-26. 10.7759/s44389-026-00042-w

How to cite this article:
Kasambya M K (November 15, 2026) Social-AgriNet: A Multimodal Vision Transformer Framework for Explainable Banana Disease Detection under Real-World Conditions. Cureus J Comput Sci 3 : es44389-026-00369-4. DOI https://doi.org/10.7759/s44389-026-00369-4