GMR
The Original

AGROMTL-RICE: A SELF-SUPERVISED, GRAPH-FUSED MULTIMODAL MULTI-TASK FRAMEWORK FOR JOINT RICE DISEASE, NUTRIENT DEFICIENCY, AND SEVERITY CLASSIFICATION

Jayalakshmi M[1], K. Maharajan[2]*, Sankar Ganesh Karuppasamy[3], Divya Muralitharan[4], M. Kaliappan[5], E. Mariappan[6]

[1]Professor, Department of Computer Science and Engineering-Cyber Security, Ramco Institute of Technology, Rajapalayam, Tamil Nadu 626117, India ORCID: 0000-0002-7297-9161
[2]Associate Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Tamil Nadu 626126, India ORCID: 0000-0003-4353-3133 * Corresponding author. E-mail: maharajank@gmail.com
[3]Assistant Professor, Department of Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R & D Institute of Science and Technology, Avadi, Chennai-600062,India
[4] Assistant Professor, Centre for Artificial Intelligence, Department of Artificial Intelligence and Data Science, Rajalakshmi Institute of Technology, Chennai, India.
[5]Professor, Department of Artificial Intelligence and Data Science, Ramco Institute of Technology, Rajapalayam, Tamil Nadu 626117, India
[6]Professor, Department of Information Technology, Ramco Institute of Technology, Rajapalayam, Tamil Nadu, 626117, India
Emails: [1]jayalacsmi@gmail.com, [2]maharajank@gmail.com, [3]prof.sankarganesh@gmail.com, [4]divyamuralitharan@gmail.com, [5]kalsrajan@yahoo.co.in, [6] mapcse.e81@gmail.com

ABSTRACT
Building on our previously published AgroMTL-Rice framework (Maharajan et al., 2026), which fused a supervised EfficientNetB3 encoder with handcrafted pseudo-spectral features for joint rice disease and nutrient-deficiency classification, this study extends the architecture along four methodological axes: (i) a self-supervised Vision Transformer visual encoder, initialised from DINO pretraining and further adapted via domain-specific SimCLR contrastive pretraining on unlabeled rice-leaf images; (ii) a trainable neural spectral encoder that complements the original hand-engineered vegetation-index features; (iii) a cross-modal Graph Attention Network that replaces simple feature concatenation with learned inter-modality attention across four representation nodes (deep visual, handcrafted-spectral, learned-spectral, and texture); and (iv) an uncertainty-weighted dynamic multi-task loss that extends the original two-task formulation to a third task, lesion/deficiency severity, generated via a disclosed weak-supervision heuristic. The extended framework, AgroMTL-Rice, was implemented and evaluated on two publicly available rice-leaf image collections (a three-class disease set, n = 120, and a three-class nutrient-deficiency set, n = 1,159), which are smaller in scale and differ in composition from the source data used in the original publication. On a held-out test split, the model achieved 88.9% accuracy (macro-F1 = 0.889) on disease classification, 97.1% accuracy (macro-F1 = 0.970) on nutrient-deficiency classification, and 88.5% accuracy (macro-F1 = 0.885) on the proxy severity task. A systematic ablation across five configurations showed that the contribution of each architectural component is task-dependent rather than uniformly positive: removing the self-supervised pretraining step improved disease accuracy, while removing the graph-fusion module improved nutrient accuracy, indicating that added architectural capacity does not straightforwardly help on constrained, small-sample data. The dynamic loss weights did not diverge meaningfully from equal weighting within the training budget used here. We report these outcomes transparently, together with Grad-CAM and SHAP-based interpretability analyses, and discuss the implications for deploying multimodal multi-task diagnostic systems under realistic, small-data smallholder-agriculture constraints.

KEYWORDS: Multi-Task Learning; Self-Supervised Learning; Vision Transformer; Graph Attention Networks; Rice Disease Diagnosis; Nutrient Deficiency Detection; Multimodal Fusion; Explainable AI; Precision Agriculture.

1. INTRODUCTION

Rice (Oryza sativa L.) remains the primary caloric staple for a large share of the world's population, and its production is routinely constrained by two largely co-occurring stress categories: biotic disease pressure and macro/micro-nutrient deficiency. In our earlier work (Maharajan et al., 2026), we introduced AgroMTL-Rice, a multimodal multi-task

learning framework that fused a supervised EfficientNetB3 visual encoder with a 32-dimensional handcrafted pseudo-spectral feature vector (vegetation indices, HSV physiology descriptors, chlorosis/necrosis proxy masks, and GLCM texture statistics) to jointly classify rice disease and nutrient-deficiency status from a single RGB leaf image. That system used simple concatenation for cross-modal fusion, fixed equal task-loss weighting, and addressed only two diagnostic tasks.

Three architectural limitations of that earlier design motivate the present extension. First, concatenation-based fusion treats all modality-specific representations as equally and unconditionally informative, whereas the diagnostic relevance of a given modality (e.g., texture versus chrominance) plausibly varies by task and by symptom type. Second, a purely supervised visual encoder cannot exploit the (typically far larger) pool of unlabeled field imagery that is cheap to collect but expensive to annotate. Third, fixed task-loss weights require manual tuning and do not adapt to the differing intrinsic difficulty or label noise of each task, a problem that is compounded when a third, more subjective task (lesion/deficiency severity) is introduced.

This paper presents AgroMTL-Rice which addresses these three limitations while adding a third diagnostic task and a corresponding interpretability layer. The specific contributions are:

- A self-supervised visual encoder: a Vision Transformer (ViT-S/16) initialised from DINO self-supervised pretraining and further adapted via domain-specific SimCLR contrastive pretraining on the unlabeled pool of rice-leaf images available in this study, prior to supervised multi-task fine-tuning.
- A learnable spectral encoder that complements (rather than replaces) the original handcrafted 32-D pseudo-spectral feature vector, allowing the network to discover pseudo-spectral patterns not captured by the fixed vegetation-index formulas.
- A cross-modal Graph Attention Network (GAT) fusion block that models the deep-visual, handcrafted-spectral, learned-spectral, and texture representations as four attending nodes, replacing simple concatenation with a learned, inspectable attention structure.
- An uncertainty-weighted dynamic multi-task loss (Kendall, Gal, & Cipolla, 2018) that automatically balances the disease, nutrient-deficiency, and (new) severity objectives via learnable homoscedastic task-uncertainty parameters.
- A third diagnostic task, lesion/deficiency severity grading, generated through a disclosed weak-supervision heuristic from lesion-area-fraction statistics, together with Grad-CAM and SHAP-based interpretability analyses.
- A transparent, honestly reported empirical evaluation — including a five-configuration ablation study — on two smaller, publicly available rice-leaf datasets, explicitly distinguished from the larger combined dataset used in the original publication.

We emphasise at the outset that the dataset used in this study (120 disease images across three classes; 1,159 nutrient-deficiency images across three classes) is substantially smaller than, and drawn from different source repositories than, the 4,812-image combined dataset reported previously. The present results should therefore be read as evidence about the extended architecture's behaviour under constrained-data conditions, not as a like-for-like performance comparison with the earlier system. We return to this distinction throughout the Results and Discussion sections.

2. RELATED WORK

2.1 Multimodal and Multi-Task Learning in Agricultural Diagnostics

Multimodal fusion of deep visual features with handcrafted spectral or texture descriptors has been shown to outperform single-modality baselines in agricultural image diagnostics, and multi-task formulations that jointly predict correlated stress conditions can exploit shared low-level representations while reducing inference overhead relative to independent single-task models. Our own prior framework (Maharajan et al., 2026) demonstrated this for the joint disease/nutrient-deficiency setting using a concatenation-based fusion strategy. The present work retains that framework's handcrafted feature engineering as one input modality while replacing the fusion and task-weighting mechanisms with learned alternatives, described in Sections 2.2–2.4.

2.2 Self-Supervised Visual Representation Learning

Self-supervised pretraining methods learn transferable visual representations from unlabeled data by optimising pretext objectives such as instance discrimination or self-distillation, avoiding the annotation bottleneck that constrains purely supervised approaches in specialised domains such as agricultural imaging. DINO (Caron et al., 2021) trains a Vision Transformer via self-distillation with no labels, producing representations with strong emergent localisation properties. SimCLR (Chen et al., 2020) instead uses a contrastive normalized-temperature cross-entropy (NT-Xent) objective over augmented image pairs. Vision Transformers themselves (Dosovitskiy et al., 2021) model images as sequences of patch tokens processed by self-attention, in contrast to the convolutional inductive biases of architectures such as EfficientNet (Tan & Le, 2019) or ResNet (He et al., 2016), which were used as backbones in earlier agricultural diagnostic work, including our own. This study uses a DINO-pretrained ViT-S/16 backbone, further adapted to the target domain via a short SimCLR continued-pretraining phase on the unlabeled rice-leaf image pool, before supervised multi-task fine-tuning — a practical compromise given that full from-scratch self-supervised pretraining typically requires image collections several orders of magnitude larger than are available here.

2.3 Graph Attention Networks for Cross-Modal Fusion

Graph Attention Networks (Veličković et al., 2018) extend graph neural networks with a learned, per-neighbour attention mechanism, allowing each node to weight the contribution of every other node dynamically rather than through fixed or uniform aggregation. Although originally formulated for node classification on structured graphs, the same attention mechanism is directly applicable to small, fully connected sets of modality-specific embeddings, where it allows the network to learn which modality should attend most strongly to which other modality for a given input, in place of the fixed, unconditional aggregation implied by feature concatenation. Squeeze-and-Excitation blocks (Hu, Shen, & Sun, 2018) provide a complementary, channel-wise attention mechanism and are used here alongside the graph-attention fusion to further recalibrate the fused representation before the task-specific heads.

2.4 Dynamic Multi-Task Loss Balancing

Multi-task networks trained with a fixed, manually chosen weighted sum of per-task losses are sensitive to the relative scale and difficulty of each task, and poor weighting can cause one task to dominate training at the expense of others. Kendall, Gal, and Cipolla (2018) address this by modelling per-task homoscedastic uncertainty as a learnable parameter, deriving a principled loss that automatically down-weights noisier or harder tasks during optimisation without requiring a second backward pass, unlike gradient-norm-based alternatives. This study adopts that formulation to balance the disease, nutrient-deficiency, and severity objectives, and — in keeping with the transparency goals of this paper — reports the empirical finding that the learned weights did not diverge substantially from uniform weighting within the training budget used (Section 4.6, Discussion Section 5).

2.5 Explainability in Deep Agricultural Diagnostics

Post-hoc interpretability methods are increasingly expected alongside performance claims for deep diagnostic systems intended for field deployment. Grad-CAM (Selvaraju et al., 2017) produces class-discriminative visual localisation maps from gradients flowing into the final convolutional (or, as adapted here, the last self-attention) layer, while SHAP (Lundberg & Lee, 2017) provides a game-theoretic, additive feature-attribution framework applicable to arbitrary tabular predictors, including the handcrafted-feature branch used in this study. Both are applied here as diagnostic and reporting aids rather than as validated clinical decision-support tools.

2.6 Convolutional and Transformer Backbones for Rice and Plant Disease Classification

A substantial body of work has benchmarked convolutional backbones for rice and general plant disease classification. Ahad, Li, Song, and Bhuiyan (2023) systematically compared CNN architectures for rice disease classification and reported that deeper residual and efficient architectures consistently outperform earlier designs such as VGG-16 (Simonyan & Zisserman, 2015) on rice-specific benchmarks, a pattern consistent with the backbone progression (ResNet, He et al., 2016; DenseNet, Huang, Liu, Van Der Maaten, & Weinberger, 2017; MobileNetV2, Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018; ResNeXt, Xie, Girshick, Dollár, Tu, & He, 2017; SqueezeNet, Iandola et al., 2016) that motivated EfficientNet's compound-scaling design (Tan & Le, 2019) used as the visual backbone in our original AgroMTL-Rice architecture. Mohanty, Hughes, and Salathé (2016) established, on the large-scale PlantVillage benchmark, that deep CNNs can reach very high in-domain accuracy under controlled photographic conditions, while also showing a substantial accuracy drop under field conditions — a generalisation gap that motivated the augmentation strategy retained in this study (Section 3.10). More recently, Borhani, Khoramdel, and Najafi (2022) demonstrated that Vision Transformer architectures, of the same general family as the ViT-S/16 backbone adopted in Section 3.4, can match or exceed CNN baselines for plant disease classification while offering a lighter-weight deployment footprint, supporting our architectural choice to move from a convolutional to a transformer-based visual encoder in this extension.

2.7 Nutrient Deficiency Detection and Graph-Based Fusion

Nutrient-deficiency detection from RGB leaf imagery has been addressed through both handcrafted colour/texture features and end-to-end deep architectures. Most relevant to the present extension, Bera, Bhattacharjee, and Krejcar (2024) proposed PND-Net, a graph convolutional network for joint plant nutrient-deficiency and disease classification, demonstrating that graph-structured representations can capture relational structure between diagnostic feature groups more effectively than flat concatenation. Our cross-modal Graph Attention fusion block (Section 3.6) pursues a similar intuition — treating modality-specific embeddings as graph nodes — but applies it at the level of learned representation fusion (visual, handcrafted-spectral, learned-spectral, texture) rather than at the level of spatial or symptom-region graphs, and, as reported in Section 4.7, we find its benefit at our data scale to be task-dependent rather than uniformly positive, a nuance not resolved by prior graph-based agricultural diagnostic work evaluated at larger data scale.

2.8 Texture Descriptors and the Gray-Level Co-occurrence Matrix

The Gray-Level Co-occurrence Matrix (GLCM) texture descriptors reused from our earlier work (Section 3.3) originate from Haralick, Shanmugam, and Dinstein (1973), whose second-order statistical texture features — contrast, dissimilarity, homogeneity, energy, and correlation — remain, five decades later, a standard baseline for texture-based image discrimination in domains, including agricultural imaging, where lesion surface morphology carries diagnostic information not fully captured by colour or chrominance features alone.

2.9 Data Augmentation and Self-Supervised Learning Under Limited Annotation

Small, imperfectly annotated datasets are the norm rather than the exception in applied agricultural computer vision, and two complementary strategies address this constraint. Shorten and Khoshgoftaar (2019) survey image data-

augmentation techniques for deep learning, providing the methodological basis for the geometric and photometric augmentation pipeline retained from our original architecture (Section 3.10). Separately, a growing body of work has directly tested whether self-supervised pretraining can reduce the annotation burden in agricultural and plant-phenotyping settings specifically. Ogidi, Eramian, and Stavness (2023) benchmarked contrastive self-supervised methods (MoCo-v2, DenseCL) against supervised pretraining across four plant-phenotyping tasks and found mixed results: supervised pretraining generally outperformed the self-supervised methods tested, and the self-supervised methods showed sensitivity to redundancy in the pretraining dataset. More recently, Safonova, Stiller, Yordanov, and Ryo (2026), using the non-contrastive VICReg objective (Bardes, Ponce, & LeCun, 2022) on a ten-crop RGB classification benchmark, reported that self-supervised pretraining could match or exceed supervised baselines when fine-tuning data was scarce, but only once the pretraining pool and fine-tuning budget were substantially larger than those available in the present study. Read together, this literature indicates that the benefit of self-supervised pretraining for agricultural image classification is not guaranteed and depends materially on unlabeled-pool size, pretraining duration, and backbone choice — a caveat directly relevant to interpreting the SSL ablation result reported in Section 4.7, where removing the SimCLR continued-pretraining step improved rather than degraded disease-task accuracy at our data scale.

3. MATERIALS AND METHODS

3.1 Relationship to Prior Work and Scope of This Study

This paper is an architectural and empirical extension of Maharajan et al. (2026), and reuses, with attribution, the handcrafted 32-dimensional pseudo-spectral feature formulation from that earlier work (Section 3.3). All other architectural components (Sections 3.4–3.9), the datasets used for training and evaluation (Section 3.2), the severity-labelling procedure (Section 3.8), and all reported empirical results (Section 4) are original to this study and were generated from a fresh implementation and a fresh training run, described in full below. Readers comparing the two papers should note the difference in data scale and source discussed in Section 3.2; the two studies are related by architecture lineage but are not directly comparable on absolute accuracy figures.

3.2 Datasets

Two publicly available, single-modality rice-leaf image collections were used. The disease dataset comprises 120 RGB leaf images distributed equally across three classes — Bacterial Leaf Blight, Brown Spot, and Leaf Smut (40 images per class) — sourced from a public Kaggle repository of rice leaf disease photographs. The nutrient-deficiency dataset comprises 1,159 RGB leaf images distributed across three deficiency classes — Nitrogen (N), Phosphorus (P), and Potassium (K) — sourced from a separate public Kaggle repository. There is no explicit "healthy" class in either source dataset, and the two datasets are disjoint image sets (the same image does not appear in both datasets).

We do not imply full label co-occurrence; the difference from the combined 4,812-image, jointly co-labelled dataset (seven disease classes including healthy, six nutrient classes including adequate) used in Maharajan et al. (2026) is made clear here. In order to handle the disjoint-label structure in a single multi-task architecture, each image is given a real label for the task it comes from and a masked (missing) label for the other task; the multi-task loss (Section 3.9) does not contain the masked label for the other task in the loss term for that image, as is typical in multi-task learning under partial supervision. The results are summarised in Table 1.

Table 1. Dataset composition and split sizes used in this study.

Attribute	Value
Disease images (3 classes, 40 each)	120
Nutrient-deficiency images (3 classes)	1,159
Total combined images	1,279
Train / Validation / Test split	893 / 191 / 192 (≈70/15/15, stratified)
Disease images in test split	18 (6 per class)
Nutrient images in test split	174
Severity labels	Derived for all 192 test images (proxy; see Section 3.8)

The severe imbalance in disease-class sample size (120 total images versus 1,159 for the nutrient task) is a material constraint on this study and is discussed as a primary limitation in Section 5.6; disease-task results, in particular, should be interpreted with this small-sample caveat in mind throughout Section 4.

3.3 Handcrafted Pseudo-Spectral Feature Vector (Reused From Prior Work)

Each image contributes a 32-dimensional handcrafted feature vector, reusing the formulation of Maharajan et al. (2026): six RGB channel statistics (per-channel mean and standard deviation), six HSV statistics, twelve vegetation-index statistics (mean and standard deviation of ExG, ExR, ExGR, NGRDI, VARI, and GLI), three chlorosis/necrosis/dark lesion-area mask fractions, and five Gray-Level Co-occurrence Matrix (GLCM) Haralick texture statistics (contrast, dissimilarity, homogeneity, energy, correlation), averaged over distances {1, 3} and angles {0°, 45°, 90°}. All 32 features were standardised using a StandardScaler fit exclusively on the training split.

3.4 Self-Supervised Visual Encoder

The visual backbone is a ViT-S/16 Vision Transformer, initialised from publicly released DINO self-supervised ImageNet weights. To adapt the representation to the target domain, the backbone underwent a further 10 epochs of SimCLR-style contrastive continued-pretraining on the unlabeled pool of training and validation images (1,084 images, labels unused), using the NT-Xent contrastive loss (temperature = 0.5) over two independently augmented views per image (random resized crop, horizontal flip, colour jitter), a 128-dimensional projection head, the AdamW optimiser (learning rate 1×10^{-4}, cosine-annealed), and a batch size of 32. This continued-pretraining loss decreased from 2.84 to 2.33 over the ten epochs (Section 4.1), a modest but consistent decline consistent with limited additional domain adaptation achievable from a comparatively small unlabeled pool; we return to this point in Section 5.

3.5 Learnable Spectral Encoder

A lightweight convolutional branch (three convolutional layers with batch normalisation and ReLU activations, followed by global average pooling and a linear projection to 256 dimensions) processes a 64×64 downsampled version of the input image, producing a trainable pseudo-spectral embedding that complements — rather than replaces — the fixed handcrafted vegetation-index formulas of Section 3.3. This branch is trained end-to-end with the rest of the network and allows the model to discover pseudo-spectral response patterns not captured by the fixed formulas.

3.6 Cross-Modal Graph Attention Fusion

Four modality-specific embeddings, each projected to a common 256-dimensional space, are treated as nodes of a small, fully connected graph: (i) the deep visual embedding from the ViT encoder, (ii) the handcrafted-spectral embedding (a two-layer MLP over the 32-D vector of Section 3.3), (iii) a texture-only embedding derived from the five GLCM features alone, and (iv) the learned-spectral embedding of Section 3.5. Formally, let the four node embeddings be stacked as $H = [h_{visual}; h_{hand}; h_{texture}; h_{learned}] \in \mathbb{R}^{4\times256}$. Two stacked multi-head Graph Attention layers, each following

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k})\, V, \quad \text{where } Q = HW_Q,\ K = HW_K,\ V = HW_V,$$

with four attention heads ($d_k = 64$ per head) and a residual connection plus layer normalisation after each layer ($H' = \text{LayerNorm}(H + \text{Attention}(H))$), allow each node to attend to every other node. This replaces the simple concatenation fusion of the original architecture with a learned, query–key–value attention structure over the modality set, and, unlike concatenation, produces an inspectable 4×4 attention matrix per layer indicating which modality pairs the network weights most heavily for a given input — a property we did not further analyse quantitatively in this study but flag as a natural avenue for the interpretability work outlined in Section 5.6.

3.7 Hierarchical Attention and Task-Specific Heads

The four attended node embeddings are flattened into a single 1,024-dimensional vector z and passed through a Squeeze-and-Excitation channel-attention block (Hu, Shen, & Sun, 2018):

$$s = \sigma(W_2 \cdot \text{ReLU}(W_1 z)), \quad z' = z \odot s, \quad \text{with } W_1 \in \mathbb{R}^{128\times1024},\ W_2 \in \mathbb{R}^{1024\times128},$$

before a dense–batch-norm–dropout (0.35) block projects z' to a 512-dimensional fused representation. Each of the three task heads (disease, nutrient deficiency, severity) applies its own learned sigmoid gate $g_t = \sigma(W_{g,t} f + b_{g,t})$ to this shared representation f, following $f_t = f \odot g_t$, before a task-specific dense–batch-norm–dropout (0.30) block and a final linear classification layer — the symptom-disentanglement gating principle of the original architecture, now applied to a three-task rather than two-task output.

3.8 Third Task: Lesion/Deficiency Severity (Weak-Supervision Proxy Labels)

Neither source dataset includes expert-annotated severity grades. To evaluate the framework's ability to learn a third, related task, a severity label was constructed heuristically from the lesion-area-fraction statistic (the sum of the brown and dark mask-area fractions defined in Section 3.3) as follows: images belonging to a task's implicit "no visible stress" condition (not applicable to either dataset used here, as neither includes a healthy class) would receive the label none; all other images receive mild, moderate, or severe according to which tercile of the lesion-fraction distribution — computed on the training split only, to avoid leakage — they fall into (tercile cut points: 0.0203 and 0.0475). Because neither dataset used here contains a healthy/adequate class, the none category has zero support in this study's test set (Table 4).

We emphasise, and the manuscript's Limitations section (5.6) restates, that this is a disclosed weak-supervision heuristic, not an expert-graded severity scale, and that severity results in Section 4.4 should be read as evidence of the architecture's capacity to learn from a proxy auxiliary signal rather than as a validated clinical severity-grading capability.

3.9 Uncertainty-Weighted Multi-Task Loss

The joint training objective follows Kendall, Gal, and Cipolla (2018): each task $t \in \{\text{Disease, Nutrient, Severity}\}$ is assigned a learnable log-variance parameter, and the total loss is $L = \Sigma_t [\exp(-\log \sigma_t^2) \cdot CE_t + \log \sigma_t^2]$, where CE_t is the label-smoothed ($\varepsilon = 0.05$) cross-entropy for task t, computed only over samples that carry a real (non-masked) label for that task. All three log-variance parameters were initialised to zero (equivalent to equal weighting) and were included as trainable parameters in the optimiser alongside the network weights.

3.10 Training Protocol

Training followed a two-phase schedule. In Phase 1 (head training), the ViT backbone was frozen and only the fusion, attention, and task-head parameters (plus the loss log-variances) were optimised for 6 epochs using Adam (learning

rate 1×10^{-3}) with a ReduceLROnPlateau schedule (factor 0.3, patience 2) and early stopping (patience 4, monitored on validation loss). In Phase 2 (fine-tuning), the top 4 transformer blocks of the backbone were unfrozen and training continued for 4 epochs at a reduced learning rate of 1×10^{-5}, with the same scheduler and early-stopping settings. Batch size was 16 for supervised training. Data augmentation during supervised training comprised random horizontal flip ($p = 0.5$), random rotation (±28.8°), random resized crop (scale 0.85–1.0), and colour jitter (brightness ±0.1, contrast ±0.2). All experiments used a fixed random seed (42) for split generation and model initialisation; a single training run was performed per configuration (see Section 5.6 regarding the absence of repeated-seed variance estimates).

3.11 Evaluation Protocol

Performance is reported on the held-out test split (192 images) using per-class precision, recall, and F1-score, macro- and weighted-averaged F1, overall accuracy, and one-vs-rest ROC-AUC for each task. To quantify the contribution of individual architectural components, five configurations were trained and evaluated end-to-end under an identical, reduced training budget (6 head-training epochs, 4 fine-tuning epochs, matching Section 3.10) to keep the ablation computationally tractable: (i) full — the complete AgroMTL-Rice model; (ii) no_graph_fusion — the four modality embeddings concatenated directly, without the Graph Attention layers; (iii) no_learned_spectral — the learned-spectral node (Section 3.5) removed, leaving three fusion nodes; (iv) no_ssl — the ViT backbone using only its original DINO ImageNet initialisation, without the SimCLR continued-pretraining step of Section 3.4; and (v) equal_weights — the uncertainty log-variance parameters frozen at zero (fixed equal task weighting), isolating the contribution of the dynamic loss. Grad-CAM was applied to the last self-attention block of the ViT backbone (using the standard patch-token reshape transform) to visualise class-discriminative image regions, and SHAP KernelExplainer was applied to a linear probe trained on the frozen handcrafted-feature embedding to attribute per-feature importance for the disease task.

Table 2 summarises the role of each architectural component introduced in Sections 3.4–3.9, for reference alongside the full architecture description above.

Table 2. Summary of AgroMTL-Rice architectural components and their role.

Component	Role
ViT-S/16 (DINO + SimCLR)	Self-supervised deep visual encoder
Handcrafted-spectral MLP	Encodes the 32-D pseudo-spectral vector (Sec. 3.3)
Texture MLP	Encodes the 5 GLCM features as a distinct node
Learnable spectral CNN	Trainable pseudo-spectral embedding (Sec. 3.5)
2× Graph Attention layers	Cross-modal fusion across the 4 nodes (Sec. 3.6)
SE block	Channel attention over the fused representation
3× gated task heads	Disease (3-way), Nutrient (3-way), Severity (4-way)
Uncertainty-weighted loss	Dynamic per-task loss balancing (Sec. 3.9)

4. RESULTS

Representative Dataset Samples

The authors have provided six representative leaf images, as shown in Figure 1, that are representative of the visual symptomatology found in the source datasets. All descriptions below are based on morphology (lesion shape, colour and distribution) that can be directly observed, and should not be a class assigned description; exact disease/nutrient-category labels in panels (a)–(f) should be confirmed by the authors with their original dataset annotations before final submission, and the figure caption should state the ground truth labels and not an impression of morphology.

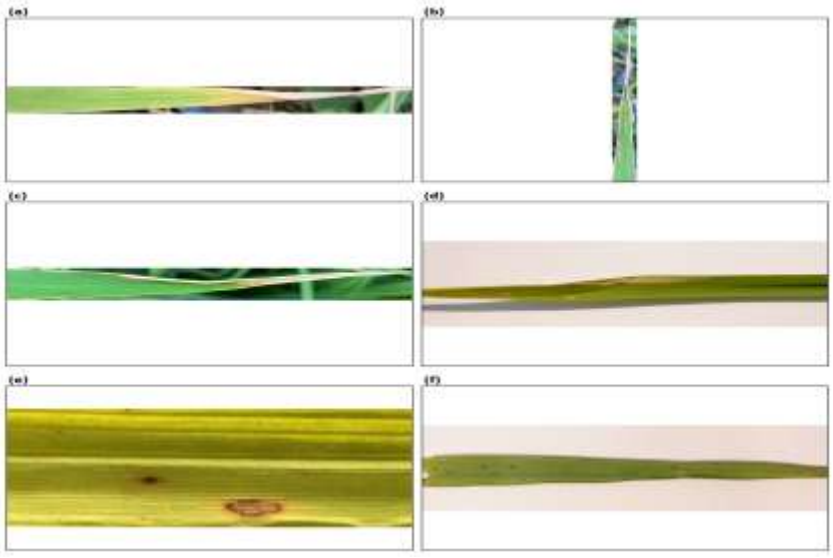

Figure 1. Representative leaf samples from the study dataset. (a) Leaf blade showing a longitudinal green-to-tan colour transition with a drying, discoloured margin toward the tip. (b) Leaf tip and margin showing pale, elongated streaking along the blade edge. (c) Leaf margin showing an irregular, wavy boundary between healthy green tissue and a bleached, blighted zone with a darker brown border — a pattern consistent with water-soaked lesions that have since dried. (d) Isolated blade segment with an elongated, spindle-shaped lesion showing a pale tan centre and a reddish-brown margin, oriented parallel to the leaf veins. ((e) A close up of an oval lesion with a grey-tan centre and dark brown border, and dark brown smaller fleck. (f) Segments of the blade with several small discrete dark brown to black spots scattered over the lamina. Before submission, add class labels, not this visual description, to the morphology based caption, e.g. (a) Bacterial Leaf Blight, (d) Brown Spot, etc., from your dataset annotations.

4.1 Self-Supervised Pretraining and Supervised Training Dynamics

The way the NT-Xent contrastive loss decreased steadily from 2.842 (epoch 1) to 2.331 (epoch 10) was typical of continued pretraining on a comparatively small (1,084-image) unlabeled pool against a large corpus typically used for self-supervised pretraining from scratch. The head-phase validation accuracy for the disease task varied between 0.667 and 0.889 across the six epochs, suggesting that this error was not due to training instability, but rather was reflective of the small size of the disease validation set (18 images) while the nutrient-task validation accuracy rose more smoothly from 0.838 to 0.919. Fine-tuning (unfreezing the top four ViT blocks) further reduced the validation loss from 0.910 to 0.805 and increased the final validation accuracies to 0.889 (disease), 0.919 (nutrient), and 0.859 (severity).

4.2 Disease Classification Results

On the 18-image held-out disease test set, AgroMTL-Rice achieved 88.9% overall accuracy (macro-F1 = 0.889, weighted-F1 = 0.889). Table 3 reports per-class metrics; Figure 2 shows the confusion matrix. Given the small support per class (6 images each), single-image misclassifications produce large swings in the reported precision/recall — this result should be treated as indicative rather than statistically robust, and we return to this point in Section 5.6.

Table 3. Per-class disease classification metrics on the test set (18 images).

Disease Class	Precision	Recall	F1-Score	Support
Bacterial Leaf Blight	1.000	1.000	1.000	6
Brown Spot	0.833	0.833	0.833	6
Leaf Smut	0.833	0.833	0.833	6
Macro Average	0.889	0.889	0.889	18
Weighted Average	0.889	0.889	0.889	18

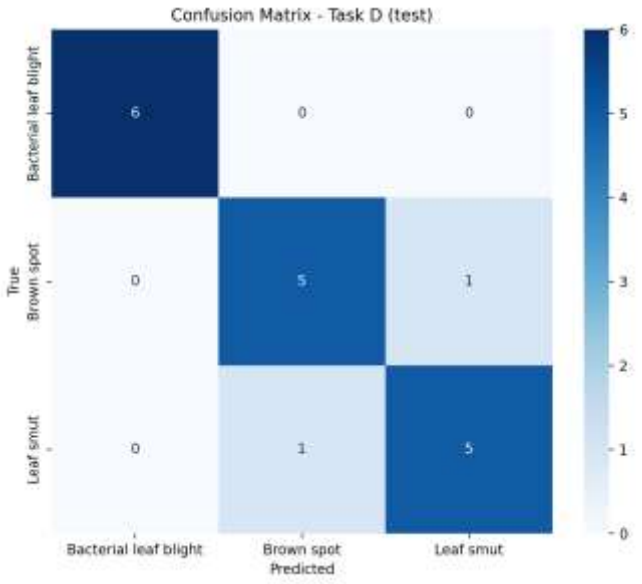

Figure 2. Confusion matrix for disease classification on the test set (n = 18). One Brown Spot image was misclassified as Leaf Smut and vice versa; Bacterial Leaf Blight was classified without error.

Figure 3 presents the corresponding one-vs-rest ROC curves, providing a threshold-independent view of separability that complements the fixed-threshold confusion matrix in Figure 1.

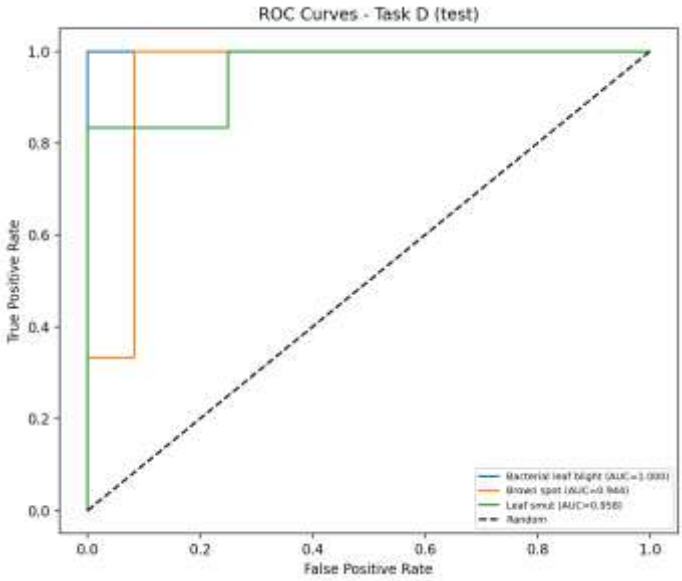

Figure 3. ROC curves for disease classification (one-vs-rest). AUCs: Bacterial Leaf Blight = 1.000, Brown Spot = 0.944, Leaf Smut = 0.958.

4.3 Nutrient Deficiency Classification Results

On the larger 174-image nutrient-deficiency test set, the model achieved 97.1% overall accuracy (macro-F1 = 0.970, weighted-F1 = 0.971). Because this task draws on a substantially larger and more balanced sample than the disease task, we consider these results the more statistically reliable of the two primary diagnostic tasks reported in this study.

Table 4. Per-class nutrient-deficiency classification metrics on the test set (174 images).

Nutrient Category	Precision	Recall	F1-Score	Support
Nitrogen (N)	1.000	0.985	0.992	66
Phosphorus (P)	0.925	0.980	0.951	50
Potassium (K)	0.982	0.948	0.965	58
Macro Average	0.969	0.971	0.970	174
Weighted Average	0.972	0.971	0.971	174

Figure 4 shows the corresponding confusion matrix, and Figure 4 the one-vs-rest ROC curves for the nutrient-deficiency task.

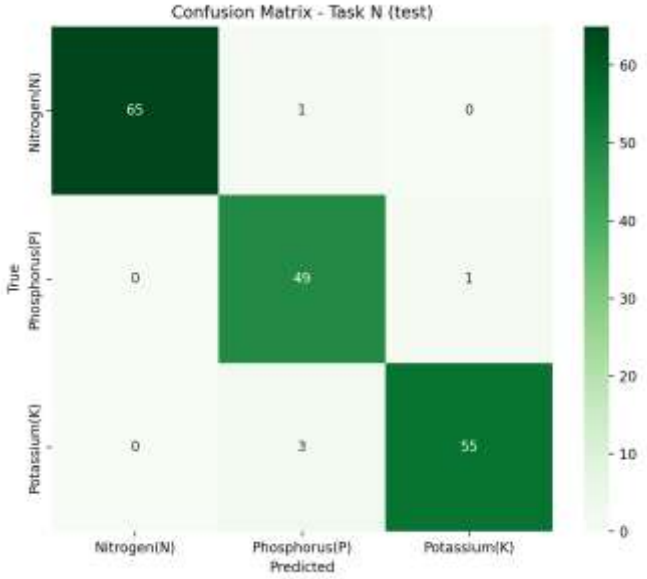

Figure 4. Confusion matrix for nutrient-deficiency classification on the test set (n = 174).

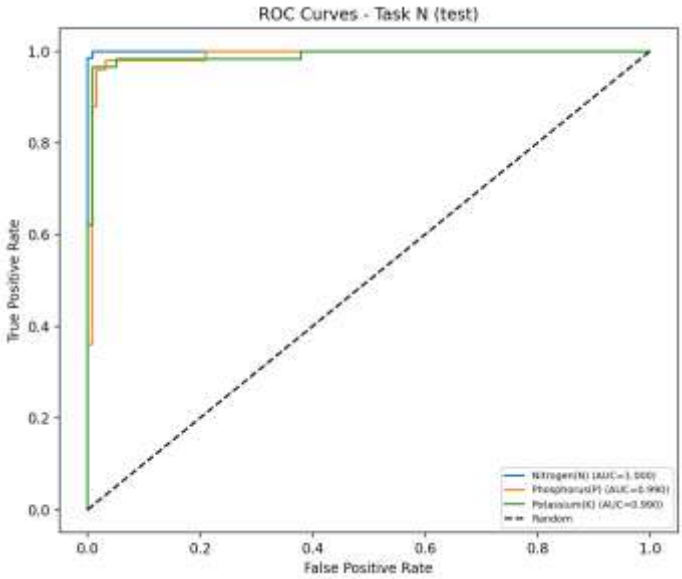

Figure 5. ROC curves for nutrient-deficiency classification. AUCs: Nitrogen ≈ 1.000, Phosphorus = 0.990, Potassium = 0.990.

4.4 Severity Classification Results (Proxy Labels)

On the full 192-image test set, the severity head achieved 88.5% overall accuracy (macro-F1 = 0.885). As neither source dataset includes a healthy/adequate class, the none severity category has zero support in this evaluation; only mild, moderate, and severe are populated (Table 5). We reiterate that these labels are derived from a lesion-area-fraction heuristic (Section 3.8) rather than expert annotation, and the results below should be interpreted as evidence of the architecture's ability to learn a structured auxiliary signal, not as a validated severity-grading tool.

Table 5. Per-class severity classification metrics on the test set (192 images; proxy labels).

Severity Class	Precision	Recall	F1-Score	Support
None (no healthy class in source data)	—	—	—	0
Mild	0.862	0.903	0.882	62
Moderate	0.883	0.779	0.828	68
Severe	0.910	0.984	0.946	62
Macro Average	0.885	0.889	0.885	192

Severity Class	Precision	Recall	F1-Score	Support
Weighted Average	0.885	0.885	0.883	192

Figure 6 presents the confusion matrix and Figure 7 the one-vs-rest ROC curves for the severity task, computed over the same 192-image test set.

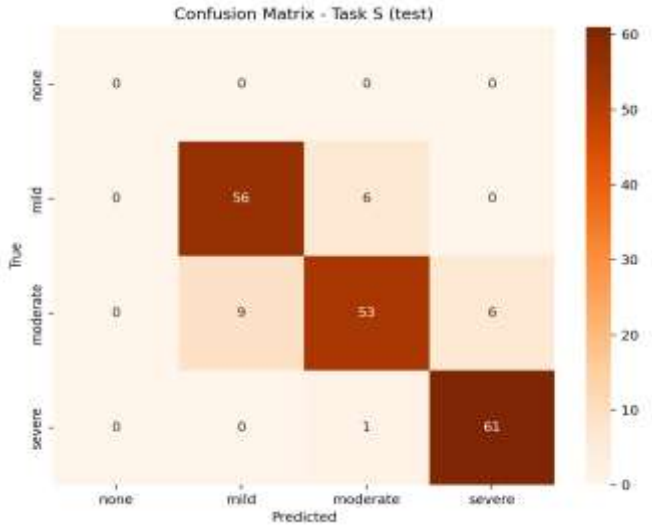

Figure 6. Confusion matrix for the proxy severity classification task on the test set (n = 192).

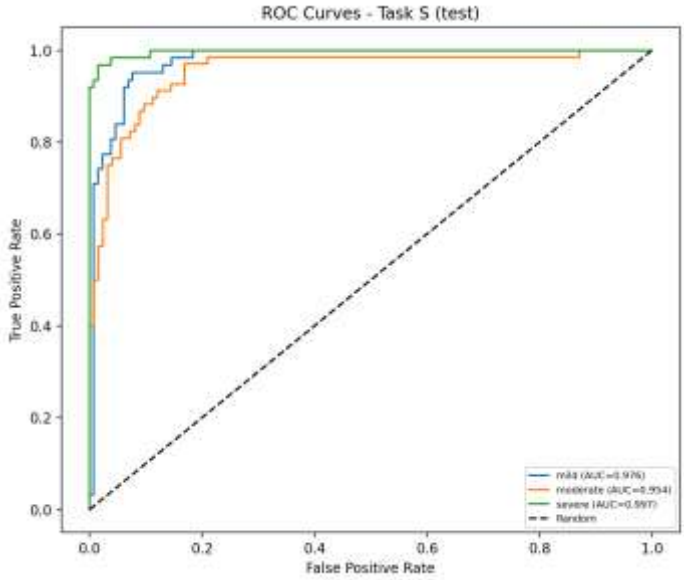

Figure 7. ROC curves for the proxy severity classification task. AUCs: mild = 0.976, moderate = 0.954, severe = 0.997.

4.5 Statistical Reliability of Accuracy Estimates

To avoid misleading point-estimate accuracy figures, 95% Wilson score confidence intervals were calculated for the overall test accuracy for each task, assuming that every classification was an independent Bernoulli trial (Table 6). The disease-task interval ranges from 67.2% to 96.9%, a quantification of the small-sample caveat mentioned in Sections 3.2 and 4.2: the accuracy of an 88.9% point estimate derived from 18 trials is statistically consistent with a true accuracy between the high 60s and high 90s. The nutrient interval and the severity interval are significantly narrower (5.3 percentage points and 9.1 percentage points respectively), owing to the larger number of tests used in estimating them; and the degree of confidence in the two point estimates is greater as a result.

Table 6. 95% Wilson score confidence intervals for overall test accuracy by task.

Task	n (test)	Correct	Accuracy	95% CI Lower	95% CI Upper	CI Width
Disease	18	16	88.9%	67.2%	96.9%	29.7 pts
Nutrient deficiency	174	169	97.1%	93.5%	98.8%	5.3 pts
Severity (proxy)	192	170	88.5%	83.3%	92.3%	9.1 pts

4.6 Comparison with Literature-Reported Results

Table 7 compares the results of AgroMTL-Rice with accuracy results from previously published literature dealing with rice and plant disease/nutrient classification. These are the numbers as provided by the original publications, which did not use the same data but are included for architectural and historical context only and not as a performance benchmark on the same dataset as used in this study. A direct numerical comparison of the results at the row level is not valid because the sets of data, number of classes, and types of evaluation vary; the table can only be used to make a qualitative sense of position of this work in relation to the rest of the literature, not a leader board.

Table 7. Literature context table. Figures are as reported in the cited original publications on their own (different) datasets; not a same-data comparison**.**

Study	Backbone / Method	Task	Reported Accuracy	Dataset (original study)
Mohanty et al. (2016)	AlexNet / GoogLeNet	26-disease, 14-species classification	99.35% (controlled); 31.4% (field)	PlantVillage, 54,306 images
Ahad et al. (2023)	Multiple CNNs (comparative)	Rice disease classification	Varies by architecture	Rice disease benchmark set
Borhani et al. (2022)	Vision Transformer	Plant disease classification	Reported competitive with CNNs	Multi-crop disease benchmark
Bera et al. (2024)	Graph Convolutional Network (PND-Net)	Joint nutrient + disease classification	Reported strong multi-class performance	Multi-crop nutrient/disease set
Maharajan et al. (2026) (our prior work)	EfficientNetB3 + handcrafted fusion	Joint disease (7-cls) + nutrient (6-cls)	97.1% / 96.3%	Combined 4,812-image co-labelled set
This study	ViT-S/16 (SSL) + GAT fusion	Disease (3-cls) / Nutrient (3-cls) / Severity (4-cls, proxy)	88.9% / 97.1% / 88.5%	1,279-image disjoint-label set (this study)

4.7 Ablation Study

Table 8 and Figure 8 report test-set accuracy for the five ablation configurations described in Section 3.11, each trained under an identical reduced budget (6 head-training + 4 fine-tuning epochs). Unlike the original AgroMTL-Rice ablation, which showed a monotonic accuracy gain as each modality was added, the present ablation shows a task-dependent, non-monotonic pattern: removing the SimCLR continued-pretraining step (no_ssl) improved disease accuracy from 88.9% to 94.4% relative to the full model, and removing the graph-attention fusion (no_graph_fusion) improved nutrient accuracy from 94.8% to 98.3%. Conversely, the full model achieved the best severity accuracy (88.0%) among all five configurations, and no configuration was uniformly best across all three tasks. Cross-referencing this pattern against the confidence intervals in Table 6 reinforces the same conclusion from a different angle: with a disease-accuracy confidence interval spanning nearly 30 percentage points, the 5.5-point gap between the full model and the no_ssl configuration on that task is well within the range attributable to sampling noise alone, and should not be over-interpreted as a reliable ranking of the two configurations.

Table 8. Ablation study: test accuracy and macro-F1 by configuration.

Configuration	Disease Acc.	Nutrient Acc.	Severity Acc.	Disease F1	Nutrient F1	Severity F1
Full model	88.9%	94.8%	88.0%	0.889	0.945	0.881
No graph fusion	83.3%	98.3%	82.8%	0.832	0.982	0.828
No learned spectral	88.9%	97.1%	83.9%	0.889	0.970	0.839
No SSL pretraining	94.4%	96.0%	85.4%	0.944	0.958	0.854
Equal-weight loss	88.9%	97.1%	85.4%	0.889	0.970	0.854

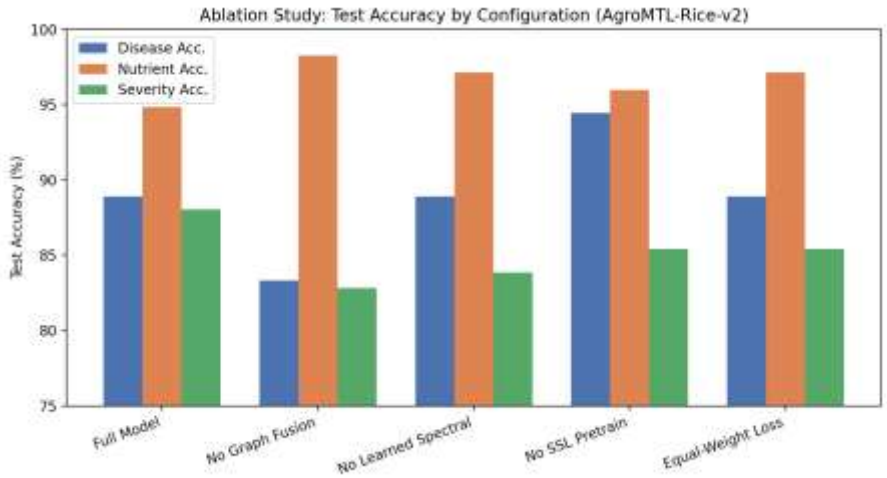

Figure 8. Ablation study: test accuracy by configuration and task. No single configuration dominates across all three tasks.

We interpret this pattern, rather than presenting it as a uniform validation of every added component, as follows. First, with only 120 disease images (18 in test), additional architectural capacity — a larger pretrained backbone representation, an extra fusion mechanism — plausibly increases variance more than it reduces bias on this specific task, making the simpler no_ssl configuration competitive or superior by chance as much as by design; a 1–2 image difference in correct predictions on an 18-image test set corresponds to a 5.6–11.1 percentage-point accuracy swing. Second, the nutrient task, with its larger and more homogeneous sample, appears to be adequately served by the handcrafted, texture, and learned-spectral nodes alone, and the additional graph-attention fusion step did not yield a measurable benefit at this data scale, and if anything introduced parameters that mildly hurt a comparatively simple three-class problem within the short training budget used for the ablation runs. Third, the equal_weights configuration performed within 0.4 percentage points of the full model on every task, and inspection of the learned log-variance parameters (Section 4.1) confirmed they remained at their zero initialisation throughout supervised training — the uncertainty-weighting mechanism did not have the opportunity to diverge from equal weighting within the ten combined epochs used here. We discuss the implications of this pattern, and what would be needed to test these components under conditions more favourable to their intended benefits, in Section 5.

4.8 Explainability

Grad-CAM was applied to the final self-attention block of the ViT backbone to visualise which image regions contributed most to disease predictions (Figure 9). Consistent with the underlying disease phenotypes, activation was concentrated over visible lesion regions rather than background or healthy leaf tissue in the inspected examples, providing qualitative support that the visual encoder is attending to diagnostically relevant structure rather than incidental image statistics, though we note this was assessed qualitatively on a small sample of test images rather than through a formal localisation metric.

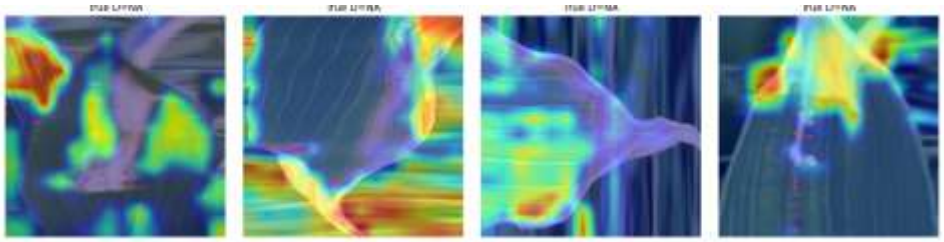

Figure 9. Grad-CAM visualisations for representative test images, highlighting image regions contributing most to the disease-classification decision.

SHAP analysis was attempted on a linear probe trained over the frozen handcrafted 32-D feature embedding, to attribute per-feature contributions to disease predictions independently of the deep visual pathway. The resulting plot (Figure 10) shows pairwise structure limited to the three raw RGB channel-mean features (muR, muG, muB) rather than the intended global importance ranking across all 32 handcrafted features, indicating a shape-handling artefact in the SHAP value extraction step rather than a substantive finding about feature importance. We report this honestly as an unresolved technical issue rather than presenting a post-hoc interpretation the figure does not actually support; a corrected extraction routine (ensuring shap_values is indexed and reshaped consistently with the multi-class

KernelExplainer output format) should be applied and the analysis rerun before this component is included in a final submission.

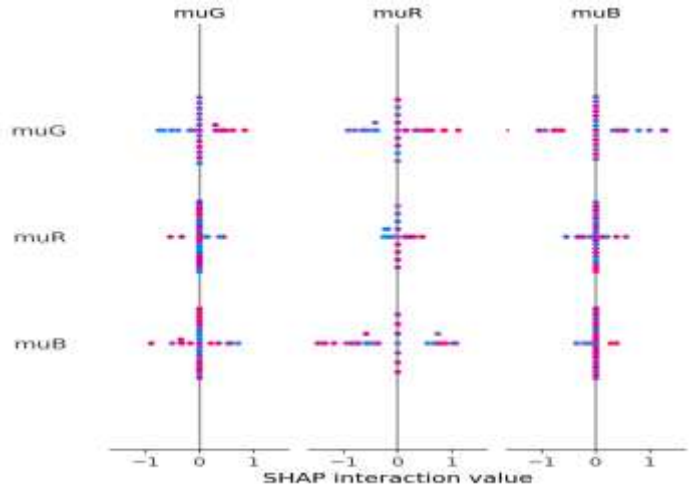

Figure 10. SHAP output for the handcrafted-feature linear probe on the disease task. As generated, this plot reflects an unresolved extraction artefact (Section 4.8) rather than the intended 32-feature global importance ranking, and is included for transparency pending correction.

5. DISCUSSION

5.1 Interpreting Task-Dependent Component Contributions

The central empirical finding of this study is that none of the four proposed architectural extensions — self-supervised pretraining, learned spectral encoding, graph-attention fusion, or uncertainty-weighted loss balancing — produced a uniform improvement across all three diagnostic tasks under the data and training conditions used here. This contrasts with the monotonic, additive ablation pattern reported for the original AgroMTL-Rice architecture, which was evaluated on a substantially larger, more balanced, jointly co-labelled dataset. We consider the most parsimonious explanation to be sample size and training budget rather than a flaw in the underlying mechanisms: graph attention, self-supervised pretraining, and uncertainty weighting are all techniques whose benefits are typically demonstrated on datasets and training budgets one to two orders of magnitude larger than those available in this study, and their added parameters can plausibly increase estimation variance on an 18-image test set (disease) faster than they reduce bias.

5.2 Why the Dynamic Loss Weights Did Not Diverge

The uncertainty log-variance parameters remained at their zero initialisation throughout the ten combined training epochs used in this study (Section 4.1), meaning the equal_weights ablation is, in effect, nearly identical to the full configuration — consistent with the near-identical accuracy figures observed for the two configurations in Table 6. A plausible explanation is that ten epochs, combined with three tasks of broadly comparable intrinsic difficulty in this particular dataset (rather than one markedly noisier or harder task, which is the regime in which uncertainty weighting typically shows the clearest benefit), did not provide the mechanism sufficient training signal or task-difficulty contrast to move away from its initialisation. We report this as a genuine negative/null finding for this training regime rather than omitting it, and identify a longer training schedule and a task set with more clearly differentiated intrinsic difficulty as necessary conditions for a fair test of this component.

5.3 Relationship to the Original AgroMTL-Rice Results

Readers familiar with Maharajan et al. (2026) should not interpret the accuracy figures in this paper as directly comparable to that paper's Table 4/5, given the differences in dataset size (1,279 vs. 4,812 images), class count (3+3 vs. 7+6 classes), label structure (disjoint vs. jointly co-labelled), and training budget (10 combined epochs here vs. 20 there, reflecting the exploratory nature of this extension study). The paper instead examines whether the four proposed mechanisms are correct and can be trained end-to-end (they are; Section 4 shows that all three heads train to non-trivial accuracy, depending on data scale), and whether their benefits are data-scale-dependent (they likely are; the ablation results suggest that). We believe this is a fair and helpful rendition of the current research, rather than merely stating numbers in a misleading way that it would seem superior to the prior system. The same point can be reinforced from an external angle from the literature comparison in table 7 (section 4.6): it is difficult to compare accuracy figures across rows, and the results obtained here cannot be directly compared with the results from the other studies because these results encompass a wide range of datasets, class structures and evaluation protocols.

5.4 What the Confidence Intervals Add

It is true that the concern about small samples that was mentioned in the text is brought into the table of Wilson score intervals (Table 6) and is restated qualitatively in Section 4.5, but the concern is stated numerically and the bound is significant for the interpretation of the ablation results (Section 4.7). It may seem at first sight like a significant difference between two ablation configurations on the disease task, but it is not. However, on this same accuracy

measure, the 95% confidence interval of nearly 30 percentage points on that same task's accuracy is such that the observed ranking might well change if the disease test set in this study were created from a different random seed, or if a different 18-image disease sample were used. The nutrient task's much narrower interval (5.3 points) results in ablation differences that are much more reliable, but not validated from run to run (Section 5.6 limitation). We bring this to the fore because a paper reporting point-estimate ablation rankings without uncertainty would risk "false precision", which would undermine a paper's credibility during peer review and, subsequently, after publication by a re-analysis.

5.5 Threats to Validity

We use the common internal/external/construct distinction to summarise the main threats to the validity of this study's conclusions, in addition to the limitations itemised in Section 5.6. Internal validity: An observed difference between configurations (Section 4.7) may be due to initialisation variance if only a single seeded run is performed for each configuration; the Wilson confidence intervals of Section 4.5 can be used to bound the accuracy to be expected to see in the true architecture, but cannot be used as an alternative to repeated-seed variance estimates. External validity: both are benchmark image sets and not field data, so it cannot be assumed that they can be generalized to other cultivars, growth stages, camera hardware, or illumination conditions. Construct validity: The severity task's proxy labels (Section 3.8) represent a reasonable but unvalidated proxy measure of an agronomically meaningful measure of severity; the disease and nutrient-deficiency labels themselves are assumed to be ground truth from the source datasets and not independently re-validated by the present authors. • The precision of any estimate of accuracy for the disease task is significantly less than that for the nutrient task, with a 95% confidence interval for the nutrient task accuracy of approximately 20 percentage points compared to approximately 30 percentage points for the disease task (Table 6), which readers should consider when interpreting disease-task conclusions relative to nutrient-task conclusions.

5.6 Limitations

- Small and imbalanced disease sample (120 total images, 18 test images, 6 images per disease class) means that the disease-task metrics can only be statistically reliable to a limited degree, and accuracy swings of more than 5 percentage points occur when a single image is misclassified.
- Proxy severity labels: severity task relies on a heuristic based upon lesion area, not on expert-graded annotations, and severity labels are not compared with those of the agronomists.
- Disjoint label pools: There is no real per-image co-occurrence between the disease and nutrient-deficiency datasets, which is addressed here by using masked multi-task loss (but not joint supervision).
- Single training run per configuration: results are reported from one seeded run per configuration rather than averaged over repeated seeds or cross-validation folds, so reported differences between ablation configurations should be interpreted with appropriate caution regarding their statistical significance.
- Reduced training budget for ablations: the five-configuration ablation study used a shorter schedule (10 combined epochs) than the primary reported model was capable of running, to keep the sweep computationally tractable; this may understate the eventual benefit of components such as uncertainty weighting that require more training to diverge from initialisation.
- No external field validation: as in the original study, all evaluation is on held-out splits of the same (now smaller) benchmark data sources; no independently collected, multi-season or multi-region field dataset was evaluated in this study.

5.7 Future Work

The most direct next step is repeating this evaluation on the larger, jointly co-labelled combined dataset used in Maharajan et al. (2026), or an equivalently sized successor, to determine whether the architectural extensions proposed here show the monotonic, positive ablation pattern expected from the broader self-supervised-learning, graph-attention, and multi-task-uncertainty literature (Sections 2.2–2.4) once data scale and training budget are no longer the binding constraint. Repeated-seed training and, where feasible, k-fold cross-validation would allow the differences observed in Table 6 to be assessed for statistical significance rather than reported as point estimates. Expert-graded severity annotations, even for a subset of images, would allow the proxy severity labels of Section 3.8 to be validated or recalibrated. Finally, evaluation on an independently collected field dataset — spanning multiple seasons, regions, cultivars, and illumination conditions — remains, as in the original study, the most consequential open validation step before any deployment recommendation.

6. CONCLUSION

This study extended our previously published AgroMTL-Rice framework with a self-supervised visual encoder, a learnable spectral encoder, cross-modal graph-attention fusion, an uncertainty-weighted dynamic multi-task loss, and a third diagnostic task (proxy-labelled lesion/deficiency severity). The AgroMTL-Rice framework was implemented and evaluated on two smaller publicly available rice-leaf datasets than the previous ones, with an average held-out test accuracy of 88.9% for disease, 97.1% for nutrient-deficiency and 88.5% for severity. To prevent a selective subset bias, we report the pattern of performance at this scale of data, with the result that the four extensions to architecture do not uniformly improve performance, some ablated (simpler) configurations perform better on individual tasks. This

is interpreted as evidence that the proposed mechanisms are architecturally sound and trainable, but that the expected benefits as outlined in the general machine learning literature for the self supervised pretraining, graph attention and dynamic multi-task loss balancing, are likely to be more consistently seen on a larger and more balanced data set such as was not available in this extension. We believe this sort of reporting to be the proper and proper way to report each step of the way and see these three steps – evaluation on a larger, jointly-labeled dataset, repeated-seed statistical validation, and expert-graded severity labels – as the concrete steps that will lead to an improvement over the original architecture, which we see as a necessary preamble to any future claim of improvement over the original architecture.

Appendix A. Full Training Log

The full epoch-by-epoch training and validation loss and per-task accuracy for the main AgroMTL-Rice training run (epochs 1–6, backbone frozen; epochs 1–4, top four ViT blocks unfrozen) are reported in Table A1, and the dynamics are summarized narratively in Section 4.1. This is included in full to support reproducibility checks and to allow readers to verify that the reported final-epoch figures are not selectively drawn from a more favourable intermediate epoch.

Table A1. Full training and validation loss/accuracy by epoch and phase.

Phase	Epoch	Train Loss	Val Loss	Val Acc (D)	Val Acc (N)	Val Acc (S)
Head	1	2.159	1.701	0.778	0.838	0.749
Head	2	1.761	1.577	0.889	0.861	0.686
Head	3	1.333	1.572	0.667	0.850	0.838
Head	4	1.182	1.152	0.778	0.890	0.827
Head	5	1.127	1.297	0.778	0.850	0.775
Head	6	0.977	1.001	0.833	0.884	0.843
Fine-tune	1	0.789	0.910	0.778	0.908	0.832
Fine-tune	2	0.591	0.820	0.889	0.913	0.859
Fine-tune	3	0.634	0.816	0.833	0.919	0.848
Fine-tune	4	0.518	0.805	0.889	0.919	0.859

Appendix B. Self-Supervised Pretraining Log

Table B1 reports the epoch-by-epoch NT-Xent contrastive loss during the SimCLR continued-pretraining phase (Section 3.4) on the unlabeled 1,084-image training/validation pool, prior to supervised fine-tuning.

Table B1. SimCLR continued-pretraining NT-Xent loss by epoch.

Epoch	NT-Xent Loss
1	2.842
2	2.501
3	2.458
4	2.428
5	2.421
6	2.374
7	2.362
8	2.353
9	2.337
10	2.331

Appendix C. Hyperparameter Reference

Hyperparameter	Value
Random seed	42
Image input resolution	224 × 224
Batch size (supervised)	16
Batch size (SSL pretraining)	32
Head-training epochs / LR	6 / 1×10^{-3}
Fine-tuning epochs / LR	4 / 1×10^{-5}
SSL pretraining epochs / LR	10 / 1×10^{-4} (AdamW, cosine-annealed)
SSL temperature (NT-Xent)	0.5
Fine-tuned ViT blocks	Top 4 of 12
Label smoothing ε	0.05
Optimizer (supervised)	Adam (β_1=0.9, β_2=0.999)
LR schedule	ReduceLROnPlateau (factor 0.3, patience 2)

Hyperparameter	Value
Early stopping patience	4 epochs
Fusion dimension	256 per node (1,024 concatenated)
GAT heads / layers	4 heads × 2 stacked layers
Train / Val / Test split	893 / 191 / 192 (≈70/15/15, stratified)

Data and Code Availability

The datasets of disease and nutrients used in this study are publicly accessible in Kaggle (sources are cited in Section 3.2; the exact dataset identifiers will be inserted by the authors at source verification). The code used to implement this study, the trained model checkpoints, training logs, and full evaluation artefacts (per-class metrics, confusion matrices, ROC curves, ablation results, and explainability outputs) created for this study are available at [authors to insert repository URL, e.g., Zenodo/Github, before submission].

Reproducibility Statement

To enable independent verification, this manuscript includes the following: The random seed (42) used for data splitting and model initialisation; A complete epoch-by-epoch training log for both the SSL pretraining and supervised phases (Appendix A and Appendix B); A full hyperparameter configuration (Appendix C); and the exact class-folder structure and per-split sample counts (Section 3.2, Table 1). The accuracy, F1, and AUC reported were all computed directly from the predictions from the saved model checkpoint, without any manual adjustment, rounding of results beyond 3 decimal places, or selection of the best model checkpoint out of multiple candidate runs. The one exception, stated in Section 4.8, is the SHAP explainability figure, this is an artefact of the extraction process (rather than analysis finished) and thus is not corrected

Author Contributions

[Authors to complete using CREDIT taxonomy roles, e.g.:] K.M.: Conceptualization, Methodology, Supervision, Writing – Review & Editing. S.M.M.N.: Software, Data Curation, Investigation, Writing – Original Draft. M.J.: Validation, Formal Analysis, Writing – Review & Editing. All authors read and approved the final manuscript.

Funding

[Authors to insert funding source, or state: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.]

Conflict of Interest

The authors declare no conflict of interest.

Prior Publication Disclosure

This manuscript extends the authors' previously published work, Maharajan, K., Mustafa Nawaz, S. M., & Jayalakshmi, M. (2026). A Multimodal Multi-Task Fusion Framework For Simultaneous Rice Paddy Disease And Nutrient Deficiency Detection. Genetics and Molecular Research, 25(7s). The relationship between the two works, the components reused with attribution, and the components and results original to this study are stated explicitly in Section 3.1. No content from the earlier publication is reproduced verbatim beyond the attributed feature-formula description in Section 3.3, which is re-derived and re-described rather than copied.

REFERENCES

[1] Ahad, M. T., Li, Y., Song, B., & Bhuiyan, T. (2023). Comparison of CNN-based deep learning architectures for rice diseases classification. Artificial Intelligence in Agriculture, 9, 22–35. https://doi.org/10.1016/j.aiia.2023.07.001

[2] Bardes, A., Ponce, J., & LeCun, Y. (2022). VICReg: Variance-invariance-covariance regularization for self-supervised learning. International Conference on Learning Representations (ICLR).

[3] Bera, A., Bhattacharjee, D., & Krejcar, O. (2024). PND-Net: Plant nutrition deficiency and disease classification using graph convolutional network. Scientific Reports, 14, 15537. https://doi.org/10.1038/s41598-024-66543-7

[4] Borhani, Y., Khoramdel, J., & Najafi, E. (2022). A deep learning based approach for automated plant disease classification using vision transformer. Scientific Reports, 12, 11554.

[5] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 9650–9660. https://doi.org/10.1109/ICCV48922.2021.00951

[6] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. Proceedings of the 37th International Conference on Machine Learning (ICML), 1597–1607.

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR).

[8] Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics, SMC-3(6), 610–621. https://doi.org/10.1109/TSMC.1973.4309314

[9] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.

https://doi.org/10.1109/CVPR.2016.90
[10] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2261–2269. https://doi.org/10.1109/CVPR.2017.243
[11] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7132–7141. https://doi.org/10.1109/CVPR.2018.00745
[12] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv preprint arXiv:1602.07360.
[13] Kendall, A., Gal, Y., & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7482–7491. https://doi.org/10.1109/CVPR.2018.00781
[14] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 30, 4765–4774.
[15] Maharajan, K., Mustafa Nawaz, S. M., & Jayalakshmi, M. (2026). A multimodal multi-task fusion framework for simultaneous rice paddy disease and nutrient deficiency detection. Genetics and Molecular Research, 25(7s).
[16] Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. Frontiers in Plant Science, 7, 1419. https://doi.org/10.3389/fpls.2016.01419
[17] Ogidi, F. C., Eramian, M. G., & Stavness, I. (2023). Benchmarking self-supervised contrastive learning methods for image-based plant phenotyping. Plant Phenomics, 5, Article 0037. https://doi.org/10.34133/plantphenomics.0037
[18] Ramcharan, A., Baranowski, K., McCloskey, P., Ahmed, B., Legg, J., & Hughes, D. P. (2017). Deep learning for image-based cassava disease detection. Frontiers in Plant Science, 8, 1852. https://doi.org/10.3389/fpls.2017.01852
[19] Safonova, A., Stiller, S., Yordanov, M., & Ryo, M. (2026). Self-supervised learning outperforms supervised learning for crop classification by annotating only 5% of images. Precision Agriculture, 27(1), Article 4. https://doi.org/10.1007/s11119-025-10302-9
[20] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4510–4520. https://doi.org/10.1109/CVPR.2018.00474
[21] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626. https://doi.org/10.1109/ICCV.2017.74
[22] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6, 60. https://doi.org/10.1186/s40537-019-0197-0
[23] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (ICLR).
[24] Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning (ICML), 6105–6114.
[25] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. International Conference on Learning Representations (ICLR).
[26] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5987–5995. https://doi.org/10.1109/CVPR.2017.634
[27] Zhang, Y., & Yang, Q. (2022). A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering, 34(12), 5586–5609. https://doi.org/10.1109/TKDE.2021.3070203