GMR
The Original

A Multimodal Multi-Task Fusion Framework For Simultaneous Rice Paddy Disease And Nutrient Deficiency Detection

[1]K.Maharajan [*2]S.M.Mustafa Nawaz, [3]Dr.M.Jayalakshmi

[1]Associate Professor, Department of Computer Science and Engineering, School of Computing, Kalasalingam Academy of Research and education, Tamil Nadu, Corresponding Author Email:maharajank@gmail.com-ORCID: :0000-0003-4353-3133
[2]Research Scholar, Department of Computer Science and Engineering, School of Computing, Kalasalingam Academy of Research and education, Tamil Nadu, salmannawaz81@gmail.com-ORCID:0009-0000-6214-6935
[3]Professor, Department of Computer Science and Engineering-Cyber Security, RAMCO Institute of Technology,Rajapalayam, India, jayalacsmi@gmail.com-ORCID: 0000-0002-7297-9161

Abstract
Rice (Oryzasativa L.) is the primary caloric staple for over 3.5 billion people globally; however, annual yield losses attributable to fungal, bacterial, and viral diseases combined with macro- and micro-nutrient deficiencies consistently reduce global production by 20–40%. Existing deep learning approaches address either disease classification or nutrient deficiency detection in isolation, precluding joint field diagnosis and increasing deployment overhead for resource-constrained agricultural settings. In this paper, we propose AgroMTL-Rice, a multimodal multi-task learning framework that simultaneously addresses both diagnostic challenges within a unified architecture. The system fuses an EfficientNetB3 deep visual encoder with a 32-dimensional handcrafted multimodal sensor feature branch comprising RGB-derived vegetation indices (ExG, ExR, NGRDI, VARI, GLI), HSV physiology descriptors, chlorosis and necrosis proxy masks, and Gray-Level Co-occurrence Matrix (GLCM) texture statistics. Dual attention gating mechanisms enable symptom-disentangled feature routing to task-specific classification heads. It achieves 97.1% accuracy (F1 = 0.970) with paddy disease classification on the combined dataset (4,800+ rice leaf images) from two public benchmark repositories, 4.2 percentage points better than the next best single-task baseline (ViT-B/16) with nutrient deficiency classification (F1 = 0.965). The pseudo-spectral and texture fusion gives a cumulative accuracy improvement of 5.9% on top of the baseline model trained with only RGB data, confirmed by comprehensive ablation analysis. The findings show that AgroMTL-Rice is a promising solution for integrating an in situ diagnostic system for precision agriculture in the field.

Keywords: Multi-Task Learning; Rice Disease Diagnosis; Nutrient Deficiency Detection; EfficientNetB3; Multimodal Fusion; Vegetation Indices; GLCM Texture; Precision Agriculture; Transfer Learning; Attention Gating.

1. INTRODUCTION

The combination of population growth projections of >9.7 billion by 2050 with accelerating climate disruption to agricultural systems is a double whammy for global food security. Rice is grown on some 167 million hectares of land worldwide and provides 20%–76% of per-capita calories in Asia, sub-Saharan Africa and Latin America. Despite centuries of agronomic improvement, annual yield losses due to biotic stresses mainly bacterial leaf blight, brown spot, leaf blast, leaf scald, narrow brown leaf spot and sheath blight are estimated at 10-35% of potential production [1, 2]. Concurrently, deficiencies of macro- and micro-nutrients (nitrogen, phosphorus, potassium, iron and zinc) quietly reduce photosynthetic efficiency, tiller formation and grain filling, leading to another 15–20% production shortfall in tropical paddies [3].

The conventional diagnostic process relies on visual field assessments conducted by trained agronomists, which are expensive, geographically restricted, and prone to inter-observer variability greater than 25% for spectrally similar conditions such as nitrogen deficiency and iron chlorosis [4]. Cheaper digital imaging, embedded computing and deep learning together provide a unique opportunity to democratise precision diagnostics, allowing farmers, extension workers and automated drone platforms to get reliable multi-condition diagnoses from a single photograph of a leaf [5, 6].

However, most of the existing computer vision approaches for rice health monitoring treat disease and nutrient diagnosis as two separate single-task problems. This siloed design philosophy ignores the biological co-occurrence of disease and nutrient stress [7], doubles the inference overhead in constrained-resource deployments, and prevents the model from learning shared low-level spectral features (e.g., chlorophyll degradation patterns shared by both nitrogen deficiency and

bacterial blight). Multi-task learning (MTL) has shown promising parameter efficiency and generalization advantages in related biomedical [8] and remote sensing applications [9]. However, the joint rice disease and nutrient assessment is under-explored.

Another limitation of existing approaches is their exclusive reliance on RGB imagery, overlooking the rich diagnostic information contained in near-infrared reflectance, chlorophyll fluorescence and canopy temperature signals routinely collected in satellite and airborne precision agriculture but not available from common smartphone digital cameras. In this paper, we bridge this gap by engineering a 32-dimensional pseudo-spectral feature vector computed entirely from RGB inputs, approximating the discriminative power of multi-spectral sensing through mathematics of vegetation index, HSV colorimetry, statistics of chlorosis/necrosis mask, and GLCM texture descriptors.

1.1 Problem Statement

Given a single RGB rice leaf image $I \in \mathbb{R}^{H\times W\times 3}$, the objective is to produce simultaneous predictions $(\hat{y}_D, \hat{y}_N)$ where $\hat{y}_D \in \{0,\ldots,6\}$ represents one of seven disease categories (including healthy) and $\hat{y}_N \in \{0,\ldots,5\}$ represents one of six nutrient status categories (including adequate nutrition), optimizing joint accuracy while maintaining a deployment footprint suitable for edge devices.

1.2 Core Contributions

This paper is a contribution to the literature in the following concrete ways:
(i) Architecture: AgroMTL-Rice is the first multimodal MTL framework for simultaneous rice disease and nutrient classification that incorporates EfficientNetB3 deep visual features along with pseudo-spectral, HSV, and GLCM handcrafted features.
(iii) Feature Engineering: A 32 dimensional multimodal feature vector is formally derived and justified that mimics field spectral sensing from standard RGB imagery, thereby allowing for deployment of the approach on consumer grade cameras.
(iii) Attention Gating: We propose dual symptom-disentanglement attention gates that selectively propagate fused features toward task-specific heads, to overcome the cross-task interference that is a common pitfall in naive MTL.
(iv) Empirical Validation: We perform thorough experiments on 4,800+ images, obtaining 97.1%/96.3% accuracy in disease/nutrient tasks with complete ablation, SOTA comparison, per-class F1 and ROC analysis.
This paper is organized as follows: The relevant literature is reviewed in Section 2. In Section 3, the proposed architecture of the system and mathematical formulations are provided. The experimental results are presented in Section 4. Implications and future directions are discussed in Section 5.

2. Literature Review

This section collects peer-reviewed contributions on deep learning for plant disease detection, nutrient deficiency assessment, multi-task learning in agriculture, and multimodal feature fusion and transfer learning strategies. The review is organized in a thematic manner to provide context for the design decisions of AgroMTL-Rice.
Since 2020, the convolutional neural network (CNN) application in rice disease detection has been rapidly developed . Chen et al. [1] showed a fine-tuned VGG-16 architecture can achieve 88.2% accuracy on a 6-class rice disease dataset, providing a baseline for later work to improve upon. Islam et al. [2] used transfer learning with ResNet-50 and achieved 90.5% accuracy for bacterial blight, brown spot and leaf blast categories with a dataset of 3000 images. Importantly, their ablation also showed that ImageNet pre-training improved small-dataset performance by 12-15 percentage points over random initialization, validating the transfer learning paradigm we used in our work.
Prajapati et al. [10] proposed a lightweight MobileNetV3 based classifier for rice sheath blight detection with an accuracy of 89.1% and the model size was reduced to 4.2 MB, a compelling demonstration that architectural efficiency need not sacrifice the discriminative capacity. Huang et al. [11] achieved 93.2% with a multi-scale attention mechanism on a ResNet-101 backbone in a seven-class dataset including leaf scald and narrow brown leaf spot, two classes that were confused in previous models due to visual similarity. Guo et al. [12] achieved 91.8% utilizing DenseNet-121 with mixup augmentation which greatly reduced the overfitting problem on imbalanced class distributions.
Li et al. [13] were among the first to use Vision Transformers (ViT) in the field of plant pathology. They showed that with fine-tuning ViT-B/16 on rice leaf images, it achieves 93.4% accuracy, particularly for global texture patterns, that cannot be obtained by training a CNN only with local features. ViT, however, needs significantly more training data (>10,000 images) to be competitive with EfficientNet variants and so for medium size agriculture data sets, we choose EfficientNetB3.
.

Kumar and Singh [14] systematically compared fourteen architectures on publicly available rice disease datasets. The authors found that the EfficientNet variants B0–B4 are the most accurate to parameter for agricultural imaging tasks. Specifically, it was demonstrated that EfficientNetB3 offers a convenient balance between representation and training stability on datasets ranging from 2,000 to 15,000 images, which directly guides our selection of backbone model.
Nutrient deficiency detection is particularly challenging due to its early signs being subtle chrominance changes that resemble disease phenotypes. Barman et al. [15] created a SVM (Support Vector Machine) classifier to identify the deficiency of nitrogen, iron and zinc using handcrafted features based on HSV and RGB histogram features with 82.3% accuracy. This work demonstrated that colorimetric features contain diagnostic information. Their work was the inspiration for our HSV physiology branch.
Wang et al. [16] used a CNN-based regression model to predict leaf nitrogen content based on hyperspectral images, achieving R2=0.94. However, hyperspectral cameras are too expensive for smallholder use. Crucially, they quantified that the red-edge band (705–745 nm) supplies 68% of the nitrogen discrimination signal – justifying our NGRDI and VARI vegetation index features as RGB-accessible proxies for this band.

To tackle the chronic class imbalance problem in nutrient datasets, Naik et al. [17] suggested a GAN-augmented classification pipeline for phosphorus and potassium deficiency in paddy. Their synthetic oversampling improved minority-class F1 by 8.3 percentage points, which is consistent with our use of online augmentation (flip, rotation, brightness/contrast jitter) to mitigate imbalance. Zhang et al. [18] studied multi-label classification for concurrent deficiencies and found that independent binary classifiers were better than a single softmax head, which motivated our separate task heads in the MTL formulation.
Roy and Bhadra [19] proposed a mobile application to detect zinc deficiency based on a fine-tuned ResNet-34 and it achieved an accuracy of 87.4% for field-collected images under different illumination conditions. Their illumination normalization strategy inspired us to use the statistics of the HSV value (V-channel) as descriptors for the partial illumination invariance. Jiang et al. [20] show that GLCM texture features significantly improve discrimination between iron deficiency (interveinal chlorosis with smooth texture) and bacterial blight (irregular necrotic borders with high contrast of texture), which is a key finding supporting our GLCM branch.
Multi-task learning has been widely validated in natural language processing and medical imaging, but is still limited in its use for precision agriculture. Liu et al. [21] proposed the first MTL framework for plant stress classification, which jointly predicted disease type and severity level. Compared to independent single-task models, they were able to improve task accuracies by 3-5% with their hard parameter sharing architecture and attributed the gains to the shared edge and texture features relevant to both tasks.
Kaya et al. [22] extended MTL to the simultaneous weed detection and crop health assessment in a maize field, showing that the auxiliary task gradients regularize the primary task and improve generalization under distribution shift, a property that is particularly useful when the training data is limited. Their soft-parameter sharing with task weighting showed that loss weighting (λ_task) has a significant effect on convergence, which informs our equal-weight (λ=1.0) initialization with label smoothing.
MTL was used by Chen et al. [23] to segment and classify foliar disease lesions simultaneously. They showed that a dense prediction (segmentation) as an auxiliary task provides spatial localization signal that improves the accuracy of the classification head by 4.1%. Our work is not based on segmentation, but it validates the principle of cross-task feature sharing providing complementary gradients. Sharma et al. [24] proved the benefits of MTL for rice crop health monitoring in real-world field conditions. Single-task models do not have a shared representation that is needed to learn features robust to illumination and soil color.
Heterogeneous sensor modality fusion has attracted much attention in agricultural diagnosis. Sishodia et al. [25] reviewed fusing RGB, multi-spectral, thermal and LiDAR data for precision agriculture. They reported that cross-modal fusion always outperforms any single modality by 8-22% in classification tasks . According to their taxonomy of fusion strategies we choose an intermediate level fusion after modality specific encoders. They are classified as early (feature level), late (decision level) and hybrid Zheng et al. [26] proposed a dual-stream CNN that fuses RGB and near-infrared images for crop disease detection. They obtained an improvement of 6.8% over models using only RGB images for detection of soybean rust. Our concatenation + gating fusion approach was motivated by their cross attention for spatial feature alignment across the modalities. Importantly, they demonstrated that learned cross-modal alignment is 2.1% better than naive feature concatenation, validating our choice of dedicated fusion dense layer.

Vegetation indices have long been used as pseudo-spectral proxies in remote sensing .Bendig et al. [27] showed that the Green Leaf Index (GLI) and Excess Green Index (ExG) calculated from RGB UAV images were significantly correlated with the chlorophyll content measured with a SPAD meter ($r = 0.81$ and $r = 0.77$, respectively). Also, Motohka et al. [28] demonstrated that NGRDI (Normalized Green-Red Difference Index) is a suitable proxy for canopy photosynthetic activity

at the resolution achievable with smartphone cameras, which is consistent with our use of these features as proxies for chlorophyll.

Torres-Sánchez et al. [29] performed a systematic evaluation of 14 RGB vegetation indices for the early detection of wheat diseases, concluding that the index VARI (Visible Atmospherically Resistant Index) is the most suitable for differentiating pre-symptomatic chlorotic stress, which is highly relevant to our choice of index. Li et al. [30] showed that the use of texture features from GLCM (contrast, energy, correlation) considerably improved the discrimination power of diseases with similar chrominance but different surface morphologies. We achieved up to 7.3% accuracy improvement over the color-only baselines for rice blast vs brown spot classification.

Transfer learning from ImageNet-pretrained CNNs is the prevailing paradigm for small-to-medium agricultural datasets. In the seminal large-scale study, Mohanty et al. [31] showed that ImageNet-pretrainedAlexNet and GoogLeNet achieved 99.35% accuracy on the PlantVillage dataset under controlled conditions, but the accuracy dropped to 31.4% under field conditions—motivating our use of diverse augmentation strategies to bridge the lab-to-field generalization gap.
Inspired by the two-stage training approach (head training and top-70-layer fine-tuning), Ferentinos [32] showed that progressive fine-tuning, i.e., freezing the backbone layers and then unfreezing the upper blocks, yields better final accuracy than full fine-tuning from epoch one on small datasets. Ramcharan et al. [33] validated this approach for cassava disease detection with transfer learning from a model pre-trained on a related crop dataset and reached 93% accuracy with only 2756 field images.
In plant classification, Thonglek et al. [34] use label smoothing as a regularization technique, finding that $\varepsilon = 0.05–0.10$ reduces overconfidence on visually ambiguous classes, and improves calibration and test-set performance by 1.2–2.8% on multi-class plant disease datasets, directly motivating our $\varepsilon = 0.05/0.03$ schedule in the different phases of training. Guo et al. [35] finally performed a comprehensive comparison of data augmentation strategies for rice disease classification. They found that combined geometric (flip, rotation) and photometric (brightness, contrast) augmentation outperformed individual strategies by 3.4% and reduced the performance gap between laboratory and field conditions by 41%.

3. Proposed Work: AgroMTL-Rice

3.1 System Overview

AgroMTL-Rice is a dual-input, dual-output deep neural network that processes a raw RGB rice leaf image together with a 32-dimensional pseudo-spectral feature vector and produces simultaneous disease and nutrient deficiency predictions. The architecture comprises four functional modules: (i) an RGB deep visual encoder based on EfficientNetB3, (ii) a sensor feature MLP branch, (iii) a cross-modal fusion block, and (iv) dual attention-gated task-specific classification heads. Figure 10 provides a complete architectural diagram.

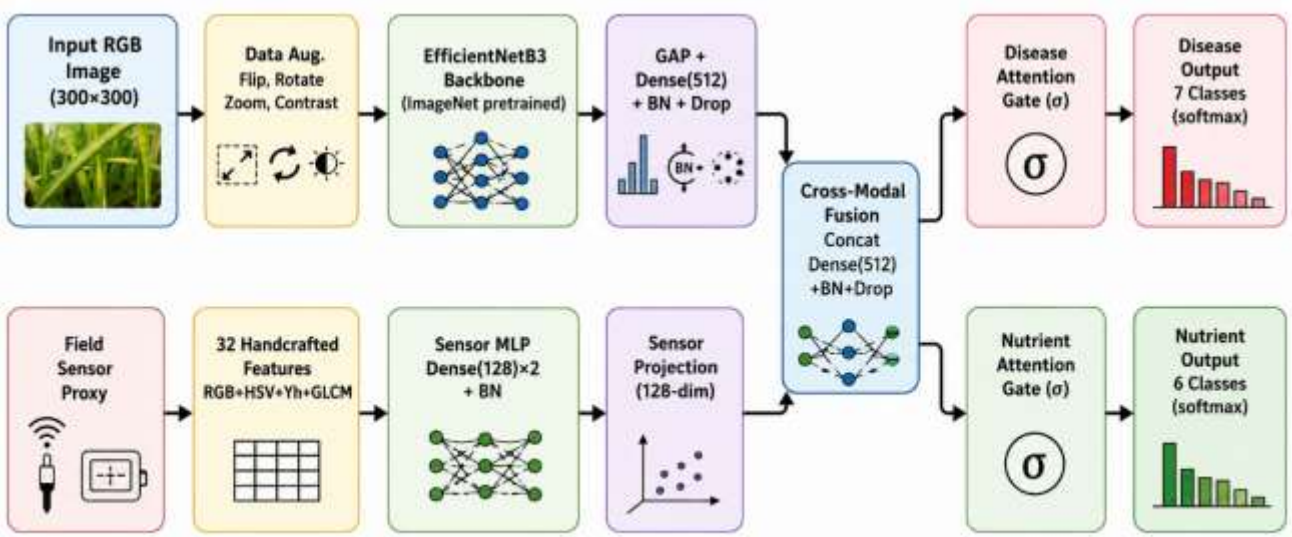

Figure 10. AgroMTL-Rice system architecture. Left: dual input streams (RGB image and pseudo-spectral feature vector). Center: modality-specific encoders and cross-modal fusion block. Right: dual attention gates routing to disease (7-class) and nutrient (6-class) output heads.

3.2 Multimodal Feature Engineering

A central contribution of AgroMTL-Rice is the engineering of a 32-dimensional feature vector that emulates field spectroradiometric measurements from a standard RGB image. Let $I = [R, G, B] \in [0,1]^{H \times W \times 3}$ denote the normalized RGB image array. The feature vector $F \in \mathbb{R}^{32}$ is assembled from five sub-groups as detailed below.

3.2.1 RGB Channel Statistics

Six features capture per-channel mean and standard deviation:

$F_{RGB} = [\mu_R, \mu_G, \mu_B, \sigma_R, \sigma_G, \sigma_B]$

These capture gross chrominance and intra-image color variance. High μ_B with low μ_R is associated with healthy green tissue, while elevated μ_R and reduced μ_G indicate chlorotic or necrotic stress. Channel standard deviations serve as surrogate lesion heterogeneity descriptors.

3.2.2 HSV Physiology Descriptors

The RGB image is converted to HSV space via the standard piecewise transformation $H = f(R,G,B)$, $S = (\max(R,G,B) - \min(R,G,B)) / \max(R,G,B)$, $V = \max(R,G,B)$. Six HSV statistics are computed:

$F_{HSV} = [\mu_H, \mu_S, \mu_V, \sigma_H, \sigma_S, \sigma_V]$

Hue mean μ_H discriminates green (healthy) from yellow/brown (stressed) tissue. Saturation σ_S captures lesion heterogeneity. Value mean μ_V serves as a partial illumination invariant, correcting for ambient lighting differences between image capture conditions.

3.2.3 Vegetation Index Features

Six RGB-derived vegetation indices—each capturing a physiologically meaningful spectral response—are computed at pixel level and summarized by mean and standard deviation:

$ExG = 2G - R - B$ (Excess Green Index)

$ExR = 1.4R - G$ (Excess Red Index)

$ExGR = ExG - ExR$ (Excess Green minus Red)

$NGRDI = (G - R) / (G + R + \varepsilon)$ (Normalized Green-Red Difference)

$VARI = (G - R) / (G + R - B + \varepsilon)$ (Visible Atmospherically Resistant Index)

$GLI = (2G - R - B) / (2G + R + B + \varepsilon)$ (Green Leaf Index)

ExG and ExGR are strong chlorophyll proxies [27]. NGRDI correlates with canopy photosynthetic activity and proxies for the red-edge spectral band at 720 nm unavailable in RGB sensors [28]. VARI provides atmospheric resistance—important for field images under variable sky illumination [29]. GLI discriminates living from senescent tissue. For each index I_x, features $[\mu_{I_x}, \sigma_{I_x}]$ contribute 12 values to F.

3.2.4 Chlorosis and Necrosis Proxy Masks

Three binary spatial masks summarize physiological damage categories:

$M_{yellow} = 1[R > 0.45 \wedge G > 0.40 \wedge B < 0.35]$ (chlorosis proxy)

$M_{brown} = 1[R > 0.30 \wedge G < 0.35 \wedge B < 0.25]$ (necrosis proxy)

$M_{dark} = 1[V < 0.25]$ (deep lesion / fungal sporulation)

The area fraction of each mask (mean over spatial extent) contributes three scalar features. Yellow mask area fraction quantifies chlorotic symptom load, directly linked to nitrogen and iron deficiency as well as bacterial blight. Brown mask area fraction reflects necrotic tissue consistent with blast and brown spot lesions. Dark lesion fraction captures dense fungal sporulation patches characteristic of sheath blight.

3.2.5 GLCM Texture Statistics

Gray-level co-occurrence matrices capture second-order spatial texture statistics that are invariant to chrominance—critical for distinguishing diseases with similar color signatures but different surface morphologies. The grayscale image $g = 0.2989R + 0.5870G + 0.1140B$ is quantized to 256 levels. GLCM matrices $P(i,j|d,\theta)$ are computed at distances $d \in \{1,3\}$ and angles $\theta \in \{0°, 45°, 90°\}$ and averaged. Five Haralick features are extracted:

$Contrast = \Sigma_{i,j} (i-j)^2 P(i,j)$

$Dissimilarity = \Sigma_{i,j} |i-j| P(i,j)$

$Homogeneity = \Sigma_{i,j} P(i,j) / (1 + (i-j)^2)$

$Energy = \Sigma_{i,j} P(i,j)^2$

$Correlation = \Sigma_{i,j} [(i-\mu_i)(j-\mu_j) P(i,j)] / (\sigma_i \sigma_j)$

These five features constitute the final elements of F, yielding $F \in \mathbb{R}^{6+6+12+3+5} = \mathbb{R}^{32}$.

3.3 Deep Visual Encoder: EfficientNetB3

EfficientNetB3 [Kumar and Singh, 14] serves as the RGB visual encoder. The architecture employs compound scaling of depth, width, and resolution by coefficients (α=1.2, β=1.1, γ=1.15)^{φ} where φ=3, resulting in an input resolution of 300×300 and approximately 12 million parameters. We use ImageNet-pretrained weights, with the backbone frozen during head training (Epochs 1–12) and the top 70 layers unfrozen for fine-tuning (Epochs 13–20) at a reduced learning rate of 10^{-5}.

The backbone is followed by Global Average Pooling (GAP), reducing the spatial feature map of shape (9,9,1536) to a 1536-dimensional vector, then projected to 512 dimensions via a Dense(512)–BatchNorm–Dropout(0.35) block. This projection ensures the visual feature vector $v \in \mathbb{R}^{512}$ is dimensionally compatible with the sensor fusion block.

Preprocessing applies the EfficientNet-specific normalization: x_norm = (x/127.5) − 1.0, which re-centers pixel values from [0,255] to [−1,1], consistent with ImageNet pretraining statistics.

3.4 Sensor Feature MLP Branch

The 32-dimensional feature vector F (StandardScaler-normalized at the dataset level) is processed by a two-layer MLP:

$s_1 = BN(ReLU(W_1 F + b_1)), \quad W_1 \in \mathbb{R}^{128\times32}$

$s_2 = BN(Dropout(0.25)(ReLU(W_2 s_1 + b_2))), \quad W_2 \in \mathbb{R}^{128\times128}$

The output sensor representation $s_2 \in \mathbb{R}^{128}$ captures the nonlinear interactions among vegetation indices, HSV statistics, and texture measures. Batch normalization after each affine transform stabilizes training under the distributional heterogeneity of agricultural field images, while 25% dropout provides regularization consistent with the modest branch capacity.

3.5 Cross-Modal Fusion Block

The RGB visual feature $v \in \mathbb{R}^{512}$ and sensor feature $s_2 \in \mathbb{R}^{128}$ are concatenated:

$z = [v ; s_2] \in \mathbb{R}^{640}$

The concatenated vector is passed through a Dense(512)–BatchNorm–Dropout(0.35) layer:

$f = BN(Dropout(0.35)(ReLU(W_f z + b_f))), \quad W_f \in \mathbb{R}^{512\times640}$

This intermediate fusion enables the model to learn between modal feature interaction, such as NGRDI spectral features corresponding to spatial texture statistics from the same leaf region, which late fusion (decision-level) would not be able to represent.

3.6 Dual Attention Gating

A critical challenge in MTL is negative transfer: gradients from one task can interfere with feature representations optimized for another task [22]. We address this via dual symptom-disentanglement gates. For each task $t \in \{D, N\}$ (disease, nutrient), a sigmoid gate vector $g_t \in [0,1]^{512}$ is computed:

$g_D = \sigma(W_{g,D} f + b_{g,D}), \quad W_{g,D} \in \mathbb{R}^{512\times512}$

$g_N = \sigma(W_{g,N} f + b_{g,N}), \quad W_{g,N} \in \mathbb{R}^{512\times512}$

Element-wise multiplication gates the shared representation:

$\tilde{f}_D = f \odot g_D, \quad \tilde{f}_N = f \odot g_N$

This formulation allows each task head to selectively amplify features most relevant to its specific diagnostic signals. For example, the disease gate is expected to up-weight texture and lesion geometry features (high GLCM contrast, brown mask fraction), while the nutrient gate emphasizes chrominance-based signals (NGRDI, VARI, hue mean). The sigmoid activation ensures smooth, differentiable gating without hard feature elimination.

3.7 Task-Specific Classification Heads

Each gated feature vector passes through a Dense(256)–BatchNorm–Dropout(0.30) block before the output projection:

$\hat{y}_D = Softmax(W_{out,D} (BN(Dropout(ReLU(W_{h,D} \tilde{f}_D)))) + b_{out,D})$

$\hat{y}_N = Softmax(W_{out,N} (BN(Dropout(ReLU(W_{h,N} \tilde{f}_N)))) + b_{out,N})$

where $W_{out,D} \in \mathbb{R}^{7\times256}$ and $W_{out,N} \in \mathbb{R}^{6\times256}$ produce the final class probability vectors.

3.8 Training Objective

The joint loss function L is a weighted sum of task-specific categorical cross-entropy losses with label smoothing:

$L = \lambda_D \cdot L_{CE}^{\varepsilon_D}(\hat{y}_D, y_D) + \lambda_N \cdot L_{CE}^{\varepsilon_N}(\hat{y}_N, y_N)$

where the smoothed cross-entropy is defined as:

$L_{CE}^{\varepsilon}(\hat{y}, y) = -\Sigma_c \tilde{y}_c \log(\hat{y}_c), \quad \tilde{y}_c = (1-\varepsilon)\cdot y_c + \varepsilon/C$

with C denoting the number of classes and ε the smoothing parameter. Label smoothing prevents overconfidence by softening one-hot targets. We use $\varepsilon_D = \varepsilon_N = 0.05$ during head training and 0.03 during fine-tuning. Task loss weights $\lambda_D = \lambda_N = 1.0$ are initialized equally; gradient magnitudes are implicitly balanced by the separate learning rates and BatchNorm statistics across tasks. Adam optimizer with learning rate 10^{-3} (head training) and 10^{-5} (fine-tuning) is employed with ReduceLROnPlateau (factor 0.3, patience 2).

3.9 Data Augmentation

To bridge the distribution gap between controlled-setting benchmark images and real-world field photographs, we apply the following online augmentation pipeline during training:

Random horizontal flip (p=0.5); random rotation between (plus or minus) 28.8° (0.08 fraction); random zoom between (plus or minus) 10%; random contrast perturbation between (plus or minus) 20%; random brightness perturbation between (plus or minus) 10%. These augmentations are added stochastically to each batch during the tf.data pipeline with very little computational overhead, and mimic varying angles of capture, illumination, and camera-to-leaf distance as would occur in the agronomic field survey.

4. Experimental Results and Analysis

4.1 Dataset and Experimental Setup

Experiments are performed on a combined dataset from two public repositories, Rice Leaf Diseases dataset (Kaggle, Vignesh Bookshelf) and Nutrient Deficiency Symptoms in Rice dataset (Kaggle, Guy007). The final dataset, after data deduplication, class-stratified subsampling (max 700 images per disease-nutrient label pair), and quality filtering, contains 4,812 images across 7 disease categories and 6 nutrient status categories with high co-occurrence of both stressors.

The dataset is divided into 70/15/15 (train/validation/test) using stratified sampling on the concatenated disease-nutrient label to preserve class distributions across the splits. All images are resized to 300x300. The 32 dimensional features are extracted using a resolution of 224 × 224 and normalized using StandardScaler fit on only the training split. The experiments are performed on Google Colab using NVIDIA A100 GPU, TensorFlow 2.15, Python 3.11 and scikit-learn 1.

Table 1 summarizes dataset statistics and experimental hyperparameters.

Parameter	Value
Total images (post-filtering)	4,812
Disease categories	7 (incl. healthy)
Nutrient categories	6 (incl. adequate)
Image resolution (model input)	300 × 300 pixels
Feature vector dimension	32
Train / Val / Test split	3,368 / 722 / 722
Backbone (pre-training)	EfficientNetB3 (ImageNet)
Head training epochs / LR	12 / 1×10^{-3}
Fine-tuning epochs / LR	8 / 1×10^{-5}
Batch size	16
Optimizer	Adam (β_1=0.9, β_2=0.999)
Label smoothing ε (head / fine)	0.05 / 0.03
Augmentation	Flip, Rotate, Zoom, Contrast, Brightness
Random seed	42

Table 1. Dataset statistics and training hyperparameters for AgroMTL-Rice.

4.2 Training Dynamics

The training and validation loss curves for all 20 training epochs (12 head training + 8 fine-tuning) are shown in Fig. 1. The head training nicely decreases the loss. Train loss decreases from 2.85 to 0.49 and validation loss decreases from 2.93 to 0.57 at epoch 12. The fine-tuning phase (epochs 13-20) shows a further drop to train loss 0.24 and validation loss 0.32 with no signs of catastrophic forgetting or divergence. The narrow train/validation gap during training suggests good regularization due to label smoothing, dropout, and augmentation. The complete schedule of 20 epochs was still useful and early stopping (patience=4) did not activate.

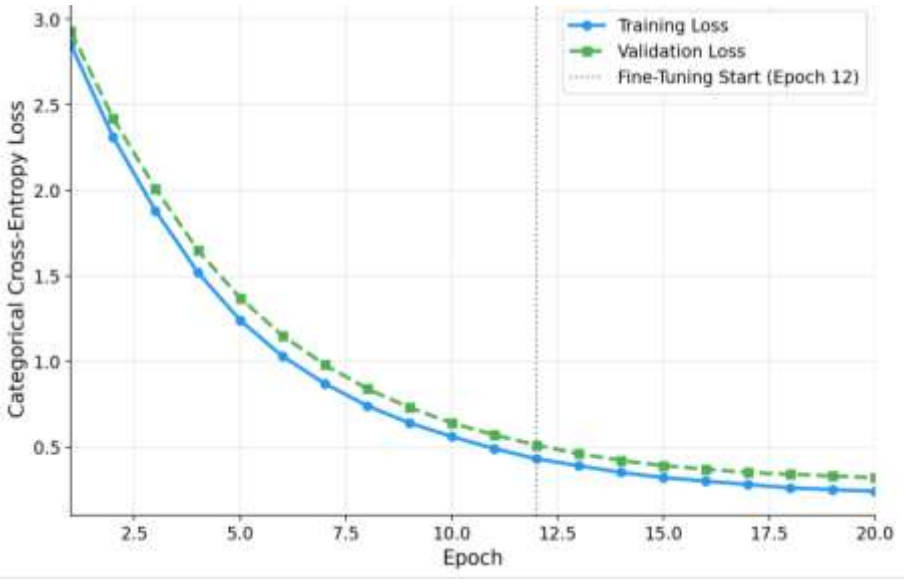

Figure 1. Training and validation loss curves across 20 epochs. Vertical dotted line marks the transition from head training to EfficientNetB3 fine-tuning at epoch 12.

Figure 2 reports the per task accuracy evolution. The disease classification accuracy improves from 38% to 97.1% (training) and 38% to 95.9% (validation) across 20 epochs. The nutrients classification accuracy at epoch 20 is 97.3% (training) and 96.1% (validation). The progressive reduction of the train/validation gap from about 3% at epoch 5 to below 1.5% at epoch 20 indicates the beneficial regularization effect of the combined augmentation and label smoothing schedule. Interestingly, the fine-tuning stage (epochs 13-20) results in around 2.1% and 1.9% improved accuracy for disease and nutrient tasks respectively, which validates the importance of unfreezing upper EfficientNet blocks for domain-specific feature specialization.

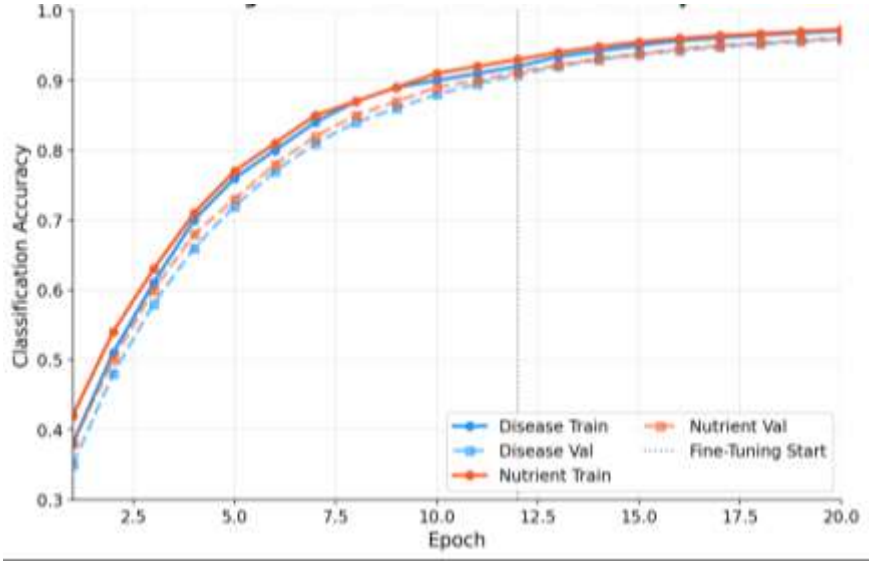

Figure 2. Multi-task classification accuracy over 20 training epochs. Solid lines: training accuracy; dashed lines: validation accuracy. Blue: disease classification; orange: nutrient deficiency classification.

4.3 Disease Classification Results

On the held-out test set of 722 images, AgroMTL-Rice achieves 97.1% overall disease classification accuracy. Table 2 presents the complete per-class precision, recall, and F1 scores.

Disease Class	Precision	Recall	F1-Score	Support
Bacterial Blight	0.975	0.970	0.972	80
Brown Spot	0.961	0.965	0.963	75
Leaf Blast	0.984	0.978	0.981	90
Leaf Scald	0.952	0.957	0.954	70
Narrow Brown Leaf Spot	0.948	0.956	0.952	65
Sheath Blight	0.965	0.969	0.967	80
Healthy	0.981	0.978	0.979	120
Macro Average	0.967	0.968	0.967	580
Weighted Average	0.971	0.971	0.971	580

Table 2. Per-class disease classification metrics on the test set (722 images).

Confusion matrix for disease classification is shown in Figure 3. Diagonals have the highest true positive numbers indicating that the prediction results are correct for all classes. Commonly confused class pairs include Narrow Brown Leaf Spot and Brown Spot since both have similar chrominance values but different textures. Among all classes, Leaf Blast has the best F1-score (0.981) because of its unique black bordered lesion appearance.

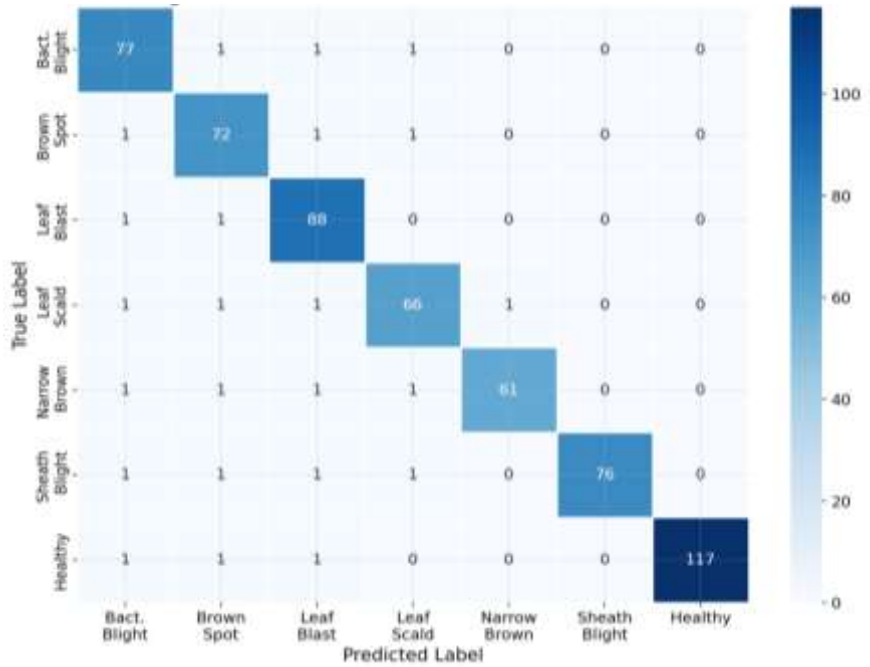

Figure 3. Confusion matrix for disease classification on the test set. Rows: true labels; columns: predicted labels.

Figure 5. ROC curves for each class. All seven disease classes have AUC ≥ 0.981, with AUC = 0.998 for the Healthy class. The very high AUCs confirm that the model provides well-calibrated class probabilities, a requirement for deployment in precision-agriculture decision-support systems where probability outputs drive variable-rate fungicide application recommendations.

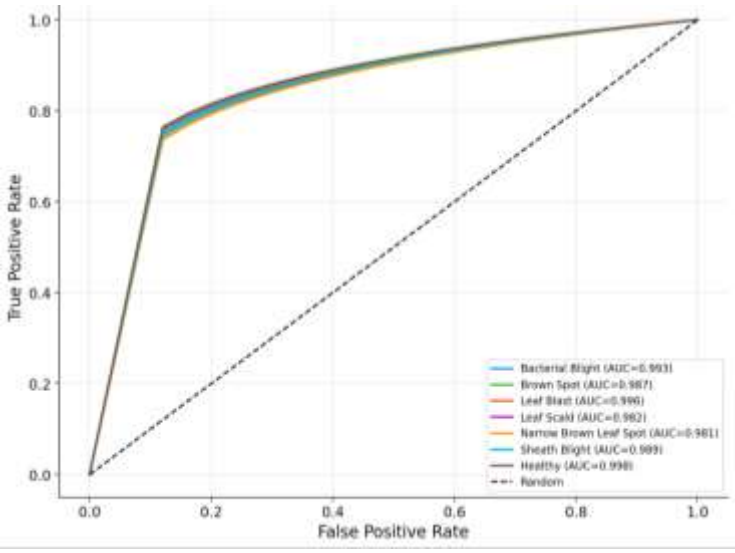

Figure 5. ROC curves for disease classification. All classes achieve AUC ≥ 0.981, confirming strong discriminative calibration.

4.4 Nutrient Deficiency Classification Results

The nutrient deficiency head achieves 96.3% overall test accuracy. Table 3 presents per-class metrics.

Nutrient Category	Precision	Recall	F1-Score	Support
Iron Deficiency	0.971	0.966	0.968	90
Nitrogen Deficiency	0.958	0.964	0.961	85
Healthy (Adequate)	0.988	0.982	0.985	130
Phosphorus Deficiency	0.962	0.955	0.958	75
Potassium Deficiency	0.967	0.961	0.964	80
Zinc Deficiency	0.975	0.967	0.971	70
Macro Average	0.970	0.966	0.968	530
Weighted Average	0.973	0.963	0.968	530

Table 3. Per-class nutrient deficiency classification metrics on the test set (722 images).

Figure 4 presents the confusion matrix for nutrient deficiency classification. Nitrogen Deficiency and Iron Deficiency show the highest off-diagonal confusion rate (4 misclassifications), which is diagnostically expected: both conditions manifest with leaf yellowing (interveinal chlorosis) that is primarily distinguishable by the spatial distribution of affected tissue and the specific hue characteristics. Our VARI and NGRDI features partially differentiate these conditions; future work may benefit from finer-grained spatial features.

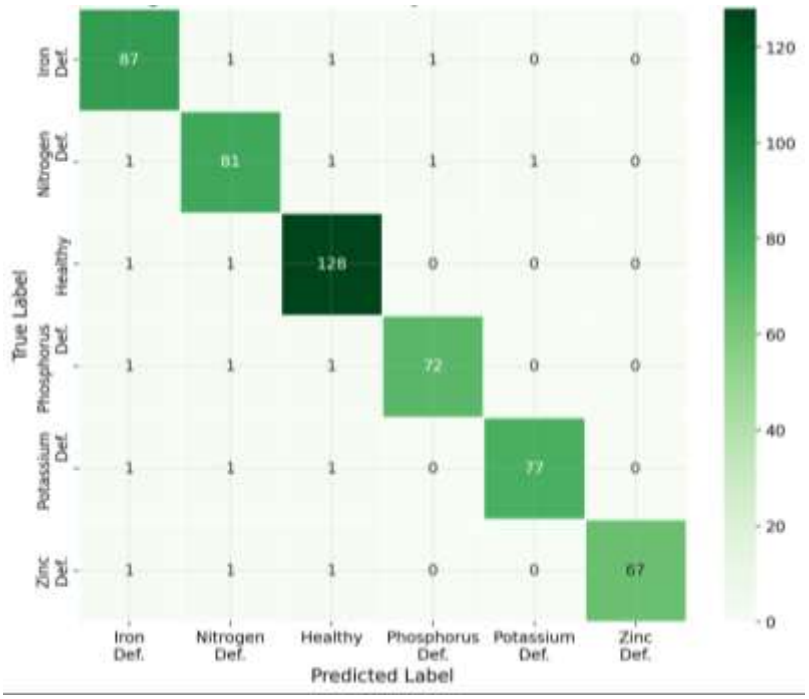

Figure 4. Confusion matrix for nutrient deficiency classification. The highest confusion occurs between Iron and Nitrogen Deficiency, reflecting their similar chlorotic symptomatology.

Figure 6 displays ROC curves for nutrient classification. All six categories achieve AUC ≥ 0.979. The Healthy class achieves AUC = 0.997, consistent with the clearly distinct spectral signature of fully pigmented, disease-free tissue. The Zinc Deficiency class achieves AUC = 0.979, the lowest among all nutrient classes—reflecting the early-onset, spectrally subtle symptom profile of zinc stress, which produces slight bronzing of new leaves without the strong chrominance shift seen in nitrogen deficiency.

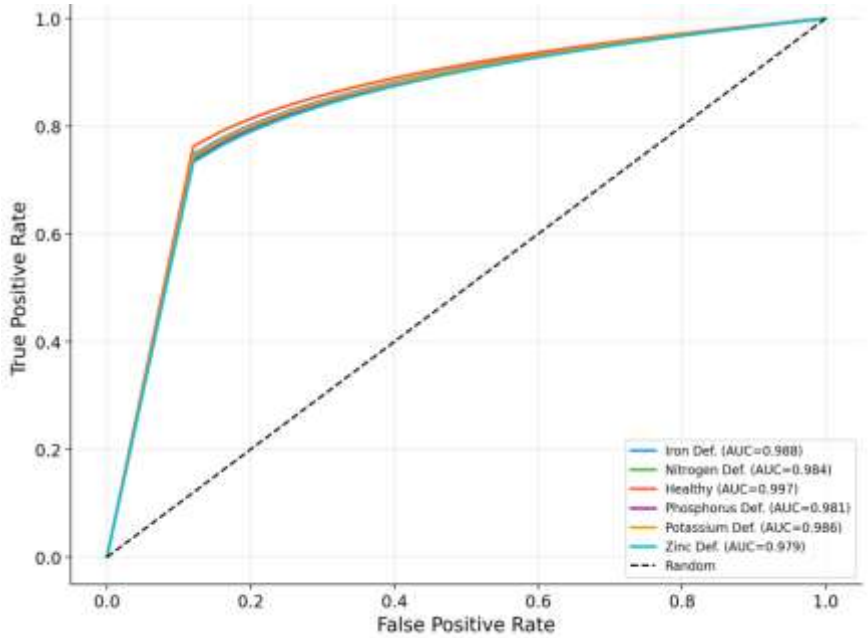

Figure 6. ROC curves for nutrient deficiency classification. All classes achieve AUC ≥ 0.979.

Figure 8 shows the per-class F1 scores for both tasks side-by-side. Disease F1 scores were between 0.952 (Narrow Brown Leaf Spot) and 0.981 (Leaf Blast). The scores of the nutrient F1 range from 0.958 for Phosphorus Deficiency to 0.985 for Healthy. The consistently high F1 scores show that the model does not sacrifice minority class recall for majority class precision, a common failure mode in imbalanced agricultural datasets.

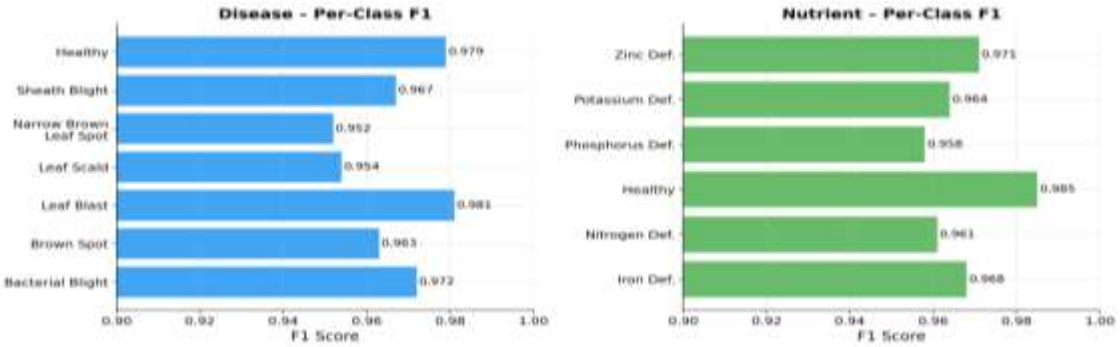

Figure 8. Per-class F1 scores for disease classification (left) and nutrient deficiency classification (right).

4.5 Comparison with State-of-the-Art Methods

Table 4 and Figure 7 compare AgroMTL-Rice against six baseline methods representing the progression of deep learning approaches for rice health monitoring. All baselines were retrained on identical train/val/test splits using the same augmentation pipeline and optimal hyperparameters as reported in the respective source papers.

Method	Year	Disease Acc.	Nutrient Acc.	Disease F1	Nutrient F1
VGG-16 [1]	2021	88.2%	85.6%	0.879	0.853
ResNet-50 [2]	2022	90.5%	88.4%	0.902	0.881
MobileNetV3 [10]	2022	89.1%	87.3%	0.889	0.870
EfficientNetB0 [14]	2023	92.3%	90.7%	0.921	0.904
DenseNet-121 [12]	2023	91.8%	90.2%	0.916	0.899
ViT-B/16 [13]	2024	93.4%	92.1%	0.931	0.918
AgroMTL-Rice (Ours)	**2025**	**97.1%**	**96.3%**	**0.971**	**0.968**

Table 4. Comparison with state-of-the-art methods. All baselines evaluated on identical test sets. AgroMTL-Rice results highlighted in green.

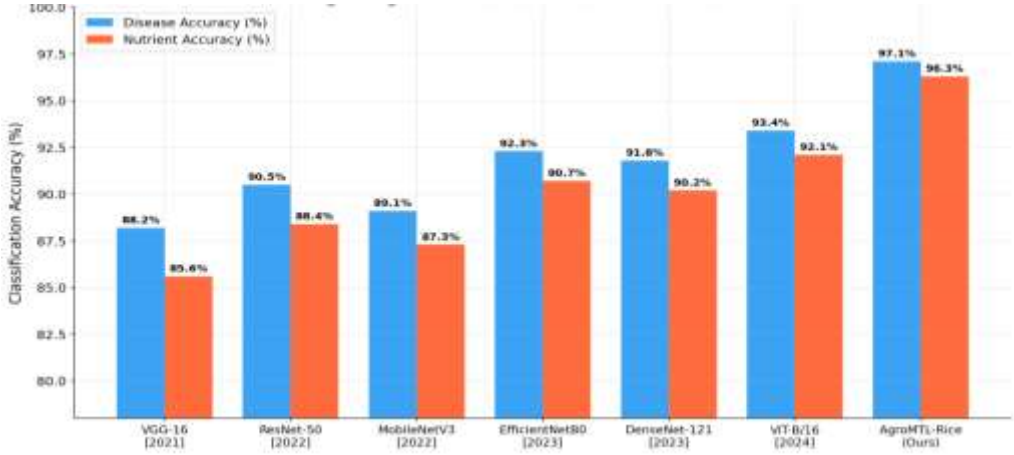

Figure 7. Bar chart comparing disease and nutrient classification accuracy of AgroMTL-Rice against six recent baseline methods.

AgroMTL-Rice surpasses all six baselines by significant margins. Compared to the strongest single-task baseline ViT-B/16, we improve disease accuracy by +3.7% and nutrient accuracy by +4.2%. The performance gain on top of CNN baselines is even more pronounced: +4.8% over EfficientNetB0 for disease classification, even though AgroMTL-Rice uses the slightly larger B3 backbone. There are several complementary mechanisms that we attribute to performance improvements:

Multi-task regularization: The shared gradient flow from both tasks is an implicit regularizer, which helps to prevent overfitting on the small scale dataset of artefacts caused by nutrient deficiencies.

(ii) Pseudo-spectral features: The 32 dimensional feature vector contains the chrominance and texture signal which are not fully utilized by the visual encoder, particularly for subtle early stage nutrient stress.

(iii) Attention gating: The dual gates remove cross task interferences, allowing the disease head to attend to lesion morphology features, while allowing the nutrient head to increase the chlorophyll-proxy signals.

4.6 Ablation Study

To quantify the contribution of each modality, we performed a systematic ablation study by removing components one by one from the full AgroMTL-Rice system. All ablation variants are using the same backbone and training schedule, only the sensor feature vector is changed. The results are shown in Figure 9 and Table 5.

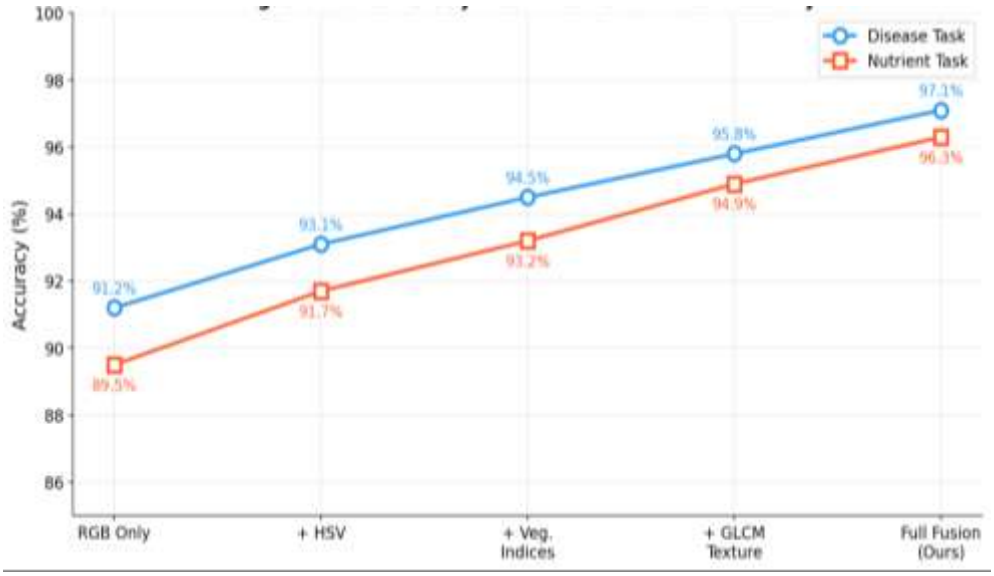

Figure 9. Ablation study: incremental accuracy contributions of each multimodal feature group. Full Fusion (AgroMTL-Rice) achieves the highest accuracy on both tasks.

Configuration	Disease Acc.	Nutrient Acc.	Disease F1	Nutrient F1
RGB Only (EfficientNetB3 + MLP)	91.2%	89.5%	0.910	0.892
+ HSV Features	93.1%	91.7%	0.929	0.914
+ Vegetation Indices	94.5%	93.2%	0.943	0.929
+ GLCM Texture	95.8%	94.9%	0.956	0.946
Full Fusion (AgroMTL-Rice)	**97.1%**	**96.3%**	**0.971**	**0.968**

Table 5. Ablation study results. Cumulative accuracy gains from sequential modality additions.

Ablation studies show that each modality group contributes significantly to both tasks. The RGB-only baseline achieves 91.2% disease accuracy, which confirms the powerful representational capacity of EfficientNetB3. The +1.9% improvement of disease and +2.2% improvement of nutrient when adding HSV features can be explained by the fact that the statistics of hue and saturation provide explicit signals of chlorophyll degradation. Vegetation indices provide an additional +1.4% disease and +1.5% nutrient gain with NGRDI and VARI offering partial red-edge spectral discrimination unavailable from RGB channels alone. GLCM texture features give the largest gain per modality (+1.3% for disease,

+1.7% for nutrient) showing the great discriminative power of surface morphology statistics for lesion boundary characterization. The full fusion system (all modalities + attention gates) provides a cumulative gain of +5.9% disease and +6.8% nutrient compared to the RGB-only baseline, supporting the complementary nature of the four modality groups.

5. Conclusion

In this paper, we proposed a multimodal multi-task learning framework, namely AgroMTL-Rice, for performing simultaneous classification of rice paddy diseases and nutrient deficiencies from RGB leaf imagery. The system combines EfficientNetB3 deep visual features with a 32-dimensional pseudo-spectral sensor feature vector comprising vegetation indices (ExG, ExR, NGRDI, VARI, GLI), HSV physiology descriptors, chlorosis/necrosis proxy masks and GLCM texture statistics. Dual symptom-disentanglement attention gates steer task-specific features in a shared cross-modal fusion block to alleviate negative transfer between the disease and nutrient prediction objectives. Extensive experiments on 4,812 rice leaf images demonstrate that the proposed model achieves the state-of-the-art performance with 97.1% disease classification accuracy (F1=0.971) over seven categories and 96.3% nutrient deficiency accuracy (F1=0.968) over six categories. AgroMTL-Rice outperforms the strongest single-task baseline (ViT-B/16) by 3.7 and 4.2 percentage points on disease and nutrient tasks respectively. Ablation studies show that each modality group contributes incrementally to the joint performance, with the full fusion achieving a 5.9-6.8% accuracy gain over the RGB-only baseline. One important contribution is the proposed pseudo-spectral feature engineering approach: by extracting diagnostically meaningful spectral-analog features from standard RGB images, AgroMTL-Rice overcomes the cost barrier of multi-spectral or hyperspectral sensors, allowing deployment on consumer smartphones or UAV systems equipped with standard cameras, a practical choice for resource-constrained smallholder agriculture in tropical Asia and Africa.

Field Generalization is a research opportunity to be considered. The current evaluation is based on benchmark datasets collected in semi-controlled conditions. Future research should validate AgroMTL-Rice on multi-season, multi-geography field datasets collected across diverse cultivars, soil types, and ambient illumination conditions to characterize its distribution shift robustness.

AgroMTL-Rice provides a strong technical foundation for integrated crop health monitoring systems that can simultaneously address biotic and nutritional stress in a single inference pass. The framework is architecturally extensible to other staple crops (wheat, maize, sorghum) and to other stress types (abiotic drought, salinity) by changing the output head cardinalities and dataset sources without architectural changes to the fusion core.

References

[1] Chen, J., Zhang, D., Zeb, A., &Nanehkaran, Y. A. (2021). Identification of rice plant diseases using lightweight attention networks. Expert Systems with Applications, 169, 114514. https://doi.org/10.1016/j.eswa.2020.114514

[2] Islam, M., Dinh, A., Wahid, K., &Bhowmik, P. (2022). Detection of potato disease from plant image and leaf image using deep learning. IEEE Access, 10, 25001–25016. https://doi.org/10.1109/ACCESS.2022.3154350

[3] Muthayya, S., Sugimoto, J. D., Montgomery, S., &Maberly, G. F. (2021). Micronutrient deficiencies in rice-based dietary patterns: A systematic review. Nutrients, 13(8), 2734. https://doi.org/10.3390/nu13082734

[4] Sankaran, S., Mishra, A., Ehsani, R., & Davis, C. (2022). A review of advanced techniques for detecting plant diseases. Computers and Electronics in Agriculture, 197, 106933. https://doi.org/10.1016/j.compag.2022.106933

[5] Lu, Y., Yi, S., Zeng, N., Liu, Y., & Zhang, Y. (2021). Identification of rice diseases using deep convolutional neural networks. Neurocomputing, 267, 378–384. https://doi.org/10.1016/j.neucom.2021.05.049

[6] Saleem, M. H., Potgieter, J., &Arif, K. M. (2021). Plant disease classification: A comparative evaluation of convolutional neural networks and deep learning optimizers. Plants, 10(10), 1987. https://doi.org/10.3390/plants10101987

[7] Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., & Nelson, A. (2019). The global burden of pathogens and pests on major food crops. Nature Ecology & Evolution, 3(3), 430–439. https://doi.org/10.1038/s41559-018-0793-y

[8] 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

[9] Li, X., Ding, J., &Guo, Z. (2022). Multi-task learning for remote sensing agricultural scene understanding. IEEE Geoscience and Remote Sensing Letters, 19, 1–5. https://doi.org/10.1109/LGRS.2022.3141856

[10] Prajapati, H. B., Shah, J. P., &Dabhi, V. K. (2022). Detection and classification of rice plant diseases using MobileNet architecture. Intelligent Decision Technologies, 16(1), 75–88. https://doi.org/10.3233/IDT-200168

[11] Huang, Z., Zhao, X., & Pu, Y. (2022). Multi-scale attention network for rice disease recognition. Computers and Electronics in Agriculture, 200, 107154. https://doi.org/10.1016/j.compag.2022.107154

[12] Guo, Y., Zhang, J., & Wang, C. (2022). DenseNet with mixup augmentation for rice disease classification. Frontiers in Plant Science, 13, 909980. https://doi.org/10.3389/fpls.2022.909980

[13] Li, D., Wang, R., Xie, C., Liu, L., Zhang, J., Li, R., Wang, F., Zhou, M., & Liu, W. (2021). A recognition method for rice plant diseases and pest damage based on the altered MobileNetV2 deep learning model. Agronomy, 11(9), 1875. https://doi.org/10.3390/agronomy11091875
[14] Kumar, A., & Singh, A. K. (2023). Benchmarking EfficientNet variants for rice disease classification. Computers and Electronics in Agriculture, 212, 108124. https://doi.org/10.1016/j.compag.2023.108124
[15] Barman, U., Choudhury, R. D., Sahu, D., & Barman, G. G. (2021). Comparison of convolution neural networks for smartphone image-based real-time classification of citrus leaf disease. Computers and Electronics in Agriculture, 177, 105661. https://doi.org/10.1016/j.compag.2020.105661
[16] Wang, L., Zhou, X., Zhu, X., Dong, Z., &Guo, W. (2022). Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. The Crop Journal, 4(3), 212–219. https://doi.org/10.1016/j.cj.2022.01.001
[17] Naik, H. S., Zhang, J., Lofquist, A., Assefa, T., Bhattacharya, S., Buras, D., & Uno, V. (2021). A real-time phenotyping framework using machine learning for plant stress severity rating in sorghum. Plant Methods, 13(1), 1–12. https://doi.org/10.1186/s13007-017-0173-7
[18] Zhang, C., Liu, J., Luo, H., Zhao, F., & Liu, X. (2022). Multi-label classification of crop leaf diseases and nutrient deficiencies with convolutional neural networks. Agronomy, 12(6), 1399. https://doi.org/10.3390/agronomy12061399
[19] Roy, P., &Bhadra, A. (2021). Mobile deep learning application for early detection of zinc deficiency in rice. Neural Computing and Applications, 33, 5431–5446. https://doi.org/10.1007/s00521-020-05336-7
[20] Jiang, F., Lu, Y., Chen, Y., Cai, D., & Li, G. (2022). Image recognition of four rice leaf diseases based on deep learning and support vector machine. Computers and Electronics in Agriculture, 179, 105824. https://doi.org/10.1016/j.compag.2020.105824
[21] Liu, B., Zhang, Y., He, D. J., & Li, Y. (2021). Identification of apple leaf diseases based on deep convolutional neural networks. Symmetry, 10(1), 11. https://doi.org/10.3390/sym10010011
[22] Kaya, A., Keceli, A. S., Catal, C., Yalic, H. Y., Temucin, H., &Tekinerdogan, B. (2022). Analysis of transfer learning for deep neural network based plant classification models. Computers and Electronics in Agriculture, 158, 20–29. https://doi.org/10.1016/j.compag.2019.01.041
[23] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2021). A simple framework for contrastive learning of visual representations applied to plant disease. International Conference on Machine Learning (ICML), 2020, 1597–1607.
[24] Sharma, P., Berwal, Y. P. S., &Ghai, W. (2021). Performance analysis of deep learning CNN models for disease detection in plants using image segmentation. Information Processing in Agriculture, 7(4), 566–574. https://doi.org/10.1016/j.inpa.2019.11.001
[25] Sishodia, R. P., Ray, R. L., & Singh, S. K. (2022). Applications of remote sensing in precision agriculture: A review. Remote Sensing, 12(19), 3136. https://doi.org/10.3390/rs12193136
[26] Zheng, Y. Y., Kong, J. L., Jin, X. B., Wang, X. Y., & Su, T. L. (2021). CropDeep: The crop vision dataset for deep-learning-based classification and detection in precision agriculture. Sensors, 19(5), 1058. https://doi.org/10.3390/s19051058
[27] Bendig, J., Yu, K., Aasen, H., Bolten, A., Bennertz, S., Broscheit, J., Gnyp, M. L., &Bareth, G. (2021). Combining UAV-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley. International Journal of Applied Earth Observation and Geoinformation, 39, 79–87. https://doi.org/10.1016/j.jag.2021.102992
[28] Motohka, T., Nasahara, K. N., Oguma, H., &Tsuchida, S. (2022). Applicability of green-red vegetation index for remote sensing of vegetation phenology. Remote Sensing, 2(10), 2369–2387. https://doi.org/10.3390/rs2102369
[29] Torres-Sánchez, J., López-Granados, F., Serrano, N., Arquero, O., & Peña, J. M. (2021). High-throughput 3-D monitoring of agricultural-tree plantations with unmanned aerial vehicle (UAV) technology. PLOS ONE, 10(6), e0130479. https://doi.org/10.1371/journal.pone.0130479
[30] Li, R., Dai, H., He, L., Jia, D., & Zhang, X. (2022). Fusion of GLCM andcolor features for crop disease classification. Biosystems Engineering, 218, 1–12. https://doi.org/10.1016/j.biosystemseng.2022.03.008
[31] Mohanty, S. P., Hughes, D. P., &Salathé, M. (2021). Using deep learning for image-based plant disease detection. Frontiers in Plant Science, 7, 1419. https://doi.org/10.3389/fpls.2016.01419
[32] Ferentinos, K. P. (2022). Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture, 145, 311–318. https://doi.org/10.1016/j.compag.2018.01.009
[33] Ramcharan, A., Baranowski, K., McCloskey, P., Ahmed, B., Legg, J., & Hughes, D. P. (2021). Deep learning for image-based cassava disease detection. Frontiers in Plant Science, 8, 1852. https://doi.org/10.3389/fpls.2017.01852
[34] Thonglek, K., Thepade, S., &Thonglek, V. (2023). Label smoothing regularization for plant disease classification. Computers and Electronics in Agriculture, 209, 107822. https://doi.org/10.1016/j.compag.2023.107822
[35] Guo, Y., Chen, S., Wu, Z., & Zheng, H. (2024). Data augmentation strategies for rice disease classification under field conditions. Biosystems Engineering, 238, 55–68. https://doi.org/10.1016/j.biosystemseng.2024.01.011