INTERNATIONAL
Agrophysics
www.international-agrophysics.org

Int. Agrophys., 2026, 40, 379-391
doi: 10.31545/intagr/2224185

Non-destructive assessment of *Hypnum* moss physiological states using a hybrid visual state space model

Zehong Lin[1]*, *Ao Fang*[2], *Minghui Cui*[2]

[1]School of Electronic Information, Lishui Vocational and Technical College, 323000, China
[2]School of Information and Control Engineering, Qingdao University of Technology, China

Received January 19, 2026; accepted May 25, 2026

A b s t r a c t. Automated quality control of *Hypnum* moss in closed bio-production is hindered by continuous phytosanitary degradation like progressive desiccation. Previous deep learning methods treat these biological states as isolated categories and rely on localized convolutions, missing dispersed morphological symptoms. This study represents the first work to apply the 2D visual state space model combined with ordinal regression to moss physiological grading. We propose the visual state space ordinal regression network (VSS-ORNet) to assess moss across four ordinal physiological grades using an expanded dataset of 1551 images. The framework synergizes a multi-branch inception stem for fine-grained textures with a 2D selective scan mechanism to capture macroscopic canopy shrinkage. A weighted ordinal loss explicitly penalizes large-margin errors to reflect the actual biophysical degradation curve. Under 5-fold cross-validation, VSS-ORNet achieved 93.49% accuracy and a 0.9662 quadratic weighted Kappa, outperforming the best baseline (Inception-ResNet-v2, 91.24%). Crucially, minimizing mean absolute error to 0.0830 ensures that misclassifications are safely confined to visually ambiguous, adjacent growth stages. Operating at 29.36 frames per second, the model satisfies latency requirements for standard batch processing. Ultimately, this non-invasive system enables real-time phytosanitary inspection, preventing cascading economic losses in large-scale bryophyte cultivation and dynamic storage facilities.

K e y w o r d s: physiological grading, visual state space models, ordinal regression, non-destructive assessment, smart agriculture

1. INTRODUCTION

Hypnum moss is a bryophyte of high commercial and ecological value, increasingly cultivated in controlled bio-production systems and smart agriculture due to its exceptional environmental resilience (Anderson *et al*., 1990; Glime, 2017). Unlike vascular plants, it utilizes a “poikilohydric” mechanism, making its physiological state-such as biomass accumulation and photosynthetic efficiency-highly sensitive to microclimate fluctuations in moisture and light (Oliver *et al*., 2005; Proctor *et al*., 2007). Consequently, continuous and precise grading of its growth stages is the primary limiting factor for quality control in large-scale production. Traditional assessment methods rely on manual sensory evaluation or destructive physical measurements (Furbank and Tester, 2011; Walter *et al*., 2015), which are labour-intensive, subjective, and fundamentally unsuitable for real-time automated control systems.

To overcome the inefficiencies of manual inspection, the agricultural sector is transitioning toward precision agriculture driven by machine vision and deep learning (Kamilaris and Prenafeta-Boldú, 2018; Liakos *et al*., 2018). While remote sensing technologies such as hyperspectral and terahertz imaging effectively capture physiological stress, their high cost and operational complexity limit

*Corresponding author e-mail: linli881010@gmail.com

widespread commercial adoption (Wakamori *et al.*, 2020). Conversely, computer vision systems based on standard RGB cameras offer a low-cost, non-contact, and real-time alternative. When integrated with advanced deep learning algorithms, RGB imaging can quantify subtle textural and chromatic variations that are imperceptible to the human eye, providing a scalable solution for continuous crop monitoring (Gulzar, 2023).

In the domain of macroscopic plant monitoring, recent literature emphasizes leveraging pre-trained vision models and transfer learning to balance classification accuracy with practical deployment. Regarding environmental stress, Hendrawan *et al.* (2021) found that ResNet50 achieved the best accuracy for grading Sunagoke moss water stress, while Anami *et al.* (2020) successfully classified paddy crop stress using field images. To ensure robustness under varying wild conditions, Picon *et al.* (2019) developed a mobile-optimized architecture for crop disease classification. More recently, the focus has shifted toward lightweight architectures and advanced transfer learning. For instance, comprehensive comparative studies on Papaya leaf disease classification have demonstrated the high efficacy of pre-trained deep learning models (Gulzar, 2025b), leading to specialized AI-driven approaches like PapNet for early detection (Gulzar, 2025c). Similarly, the application of transfer learning has been significantly advanced in tasks such as sunflower disease detection, addressing diverse environmental challenges (Gulzar, 2025a). Collectively, these macroscopic studies excel at capturing global plant states but often lack the architectural design needed to resolve fine-grained degradation.

While macroscopic models focus on overall plant health, a distinct and fundamentally different line of research addresses fine-grained image classification and microscopic texture analysis. Capturing subtle morphological details is paramount in this context. For instance, Gündüz and Günal (2024) introduced DiatomNet for classifying microscopic diatom textures with high accuracy, and Milis *et al.* (2025) applied deep learning to automate moss species identification from spore morphology. To handle complex structural variations, PTL-Inception effectively integrated deep learning with taxonomy for desert plant classification (Gulzar *et al.*, 2025). However, applying CNNs to fine-grained agricultural tasks often encounters severe spatial scale loss, a challenge recently mitigated in Date Fruit grading by explicitly preserving hierarchical features (Haque *et al.*, 2025). Furthermore, unlike general crop classification, fine-grained biological grading is notoriously plagued by data scarcity. The impact of dataset quality on deep learning models is highly pronounced in specialized tasks like dragon fruit and leaf health classification (Ayoub *et al.*, 2025). To combat the overfitting risks associated with limited samples (such as our dataset of 646 images), researchers are increasingly exploring generative techniques like diffusion models to synthesize high-quality agricultural images (Zangana *et al.*, 2025). Despite these advancements, current single-stream CNNs remain limited. They struggle to simultaneously capture the global coverage for macroscopic grading and the fine-grained textures for microscopic analysis under data-scarce conditions. This limitation necessitates the novel hybrid approach proposed in this study.

The primary aim of this work is to establish a robust, automated grading system for moss suitable for deployment in smart agriculture systems. To achieve this, we propose the visual state space ordinal regression network (VSS-ORNet), a hybrid framework explicitly designed to address the spatial and data-efficiency limitations of conventional CNNs. The main contributions of this study are summarized as follows:

– Novel Hybrid Architecture for Spatial Continuity: We design VSS-ORNet, which synergizes a lightweight, multi-branch inception stem with advanced VMamba blocks. The inception stem preserves fine-grained textural fide-lity from the earliest layers, while the 2D selective scan (SS2D) mechanism seamlessly captures global macroscopic dependencies with linear computational complexity, avoiding the prohibitive overhead of standard self-attention.

– Tailored Optimization for Ordinal Biological Data: To address the specific statistical challenges of moss health assessment, we integrate a weighted ordinal loss. This tailored optimization strictly enforces rank consistency to respect the continuous metric nature of physiological degradation, while simultaneously mitigating the inherent class imbalance present in our highly constrained dataset.

– State-of-the-Art Performance and Interpretability: Through comprehensive ablation studies and rigorous benchmarking, we demonstrate that VSS-ORNet significantly outperforms seven strong baselines spanning both CNN and transformer paradigms, including VGG16 (Simonyan and Zisserman, 2014), DenseNet121 (Huang *et al.*, 2017), Inception-ResNet-v2 (Szegedy *et al.*, 2017), EfficientNet-B4 (Tan and Le, 2019), ConvNeXt (Liu *et al.*, 2022), ViT (Dosovitskiy, 2020), and Swin-T (Liu *et al.*, 2021). Furthermore, VSS-ORNet generates highly interpretable activation maps that effectively and logically cover the entire plant area rather than spurious background noise.

Ultimately, this research demonstrates that strategically combining local texture extraction with global state-space modelling provides a highly reliable, non-invasive metric for quantifying plant health in closed bio-production systems.

2. MATERIALS AND METHODS

2.1. Source of the moss dataset and photography conditions

The dataset comprises 1 551 high-resolution images of *Hypnum* moss, classified into four ordinal growth grades based on coverage, activation status, and pigmentation (Table 1). Grade 1 represents the optimal state (401 images),

Table 1. Physiological classification criteria and dataset distribution for *Hypnum* moss grades

Grade	Coverage	Status	Colour	Dataset
Grade 1	High (>90%)	Fully active	Vibrant green	401
Grade 2	Moderate (70-90%)	Mostly active	Green	488
Grade 3	Low-moderate (40-70%)	Partially inactive	Dull Green	392
Grade 4	Very low (<40%)	Inactive	Brown	270
		Total		1 551

characterized by >90% coverage and vibrant green gametophytes. Grade 2 (488 images) denotes a good state with 70-90% coverage and slight stress, while Grade 3 (392 images) reflects a stressed state with 40-70% coverage and visible browning. Grade 4 (270 images) signifies the in-active state, showing <40% coverage and dormancy. To capture this continuous physiological degradation, the moss samples were initially cultured for eight weeks under tightly controlled indoor conditions, maintained at a temperature of 25±1°C, a relative humidity of 60-80%, and a diffused light intensity of approximately 120 μmol m^{-2} s^{-1}. Finally, to eliminate subjective bias and ensure diagnostic integrity, the preliminary evaluations of the ordinal ground-truth annotations were rigorously checked and cross-validated by two agricultural experts and two computer experts.

Images were acquired under the same conditions using a Canon EOS 90D DSLR camera with an EF-S 18-135 mm lens under controlled diffused LED lighting (5 000 K) to minimize artifacts (Fig. 1). The camera was positioned 40 cm above the samples at a 90° top-down angle. Acquisition parameters were standardized to an aperture of f/5.6, exposure of 1/100 s, ISO 2,000, and focal length of 27 mm. Original images were uniformly resized to 224×224 pixels for model input. To ensure rigorous evaluation and prevent data leakage, rather than employing a simple random split, the dataset was strictly partitioned at the physical cultivation tray level using a stratified shuffle split strategy. This resulted in a training set (70%, 1 085 images) and a testing set (30%, 466 images) for each fold of the subsequent cross-validation process.

2.2. Visual state space model with 2d selective scanning mechanism

The overall architecture of our proposed hybrid framework, termed VSS-ORNet, is illustrated in Fig. 2. To achieve this without overwhelmingly increasing model complexity on a limited dataset, we synergize the local inductive bias of a lightweight convolutional neural network with the long-range, linear-complexity modeling capabilities of the visual state space model. The following subsections detail how these components are theoretically adapted to the 2D image domain and efficiently implemented. The following subsections detail the theoretical foundations and implementation of these core components.

2.2.1. Preliminaries of state space models

The foundation of our approach lies in the continuous-time state space model, which maps an input stimulation $u(t)$ to an output response $y(t)$ through a latent state $h(t)$. Mathematically, this system is governed by the linear ordinary differential equation formulated by Gu *et al.* (2021). To adapt this to the spatial domain of moss images, let the feature map extracted by the CNN backbone be denoted as $X \epsilon \mathbb{R}^{H \times W \times C}$. We flatten spatial dimensions to treat the image as a sequence of length $L = H \times W$, such that the input at spatial position $u(t) \epsilon \mathbb{R}^{C}$. The system is governed by the linear ordinary differential equation:

a) b) c) d)

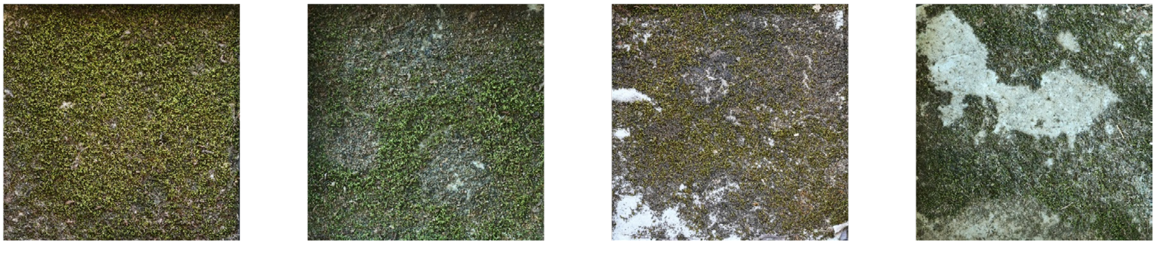


Fig. 1. Visual representation of the four moss growth grades: a) Grade 1: fully hydrated and green; b) Grade 2: mild yellowing; c) Grade 3: significant browning and curling; d) Grade 4: desiccated and dormant.

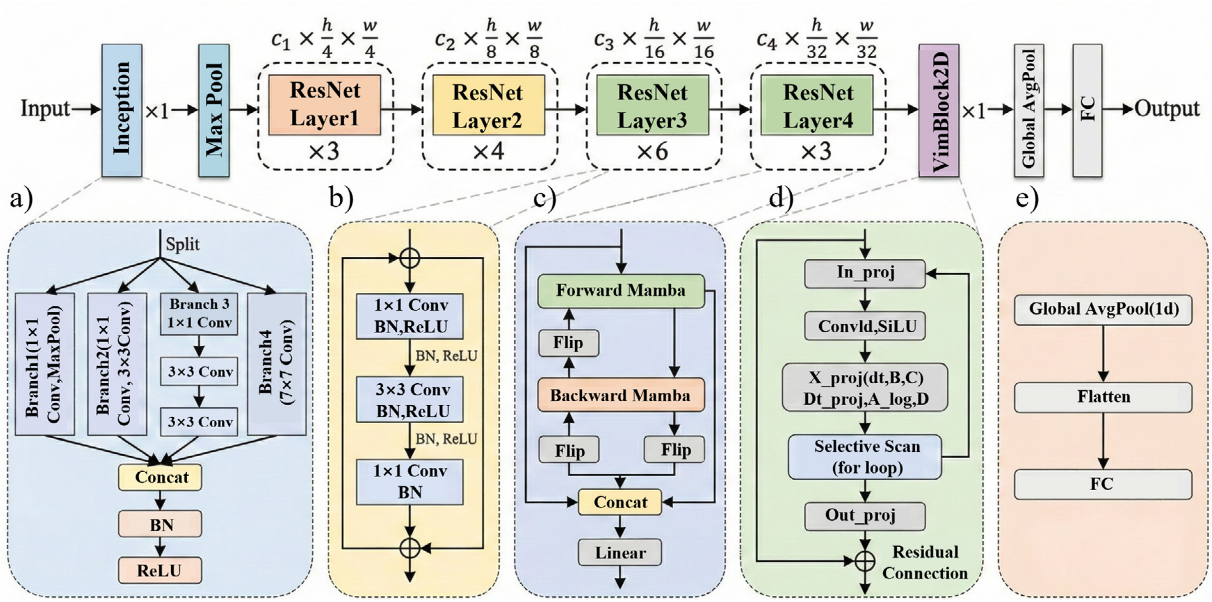


Fig. 2. Overview of the VSS-ORNet architecture: a) inception stem: initial multi-branch block capturing diverse spatial frequencies from microscopic spores to macroscopic patterns; b) ResNet bottleneck: the core stacking unit for hierarchical feature extraction; c) VimBlock2D: a bidirectional module that traverses features in forward and backward directions to model global context; d) mamba layer: the internal mechanism of the SSM, employing linear projections and selective scans to achieve linear complexity; e) classification head: maps the globally refined features to ordinal health grades.

$$\boldsymbol{h}'(t)=\boldsymbol{A}\boldsymbol{h}(t)+\boldsymbol{B}u(t), \tag{1}$$

$$y(t)=\boldsymbol{C}\boldsymbol{h}(t)+D\boldsymbol{u}(t), \tag{2}$$

where: $A \epsilon R^{N \times N}$ represents the evolution matrix, and B, C, and D are the weighting parameters. To implement this in a deep learning framework for discrete image data, we apply the zero-order hold method for discretization (Tustin, 1947). By introducing a timescale parameter Δ, the continuous parameters are transformed into their discrete counterparts $\overline{A}$ and $\overline{B}$. This allows the system to be computed recurrently:

$$\boldsymbol{h}_t=\overline{\boldsymbol{A}}h_{t-1}+\overline{\boldsymbol{B}}\,x_t, \tag{3}$$

$$y_t=\boldsymbol{C}\boldsymbol{h}(t)+D\boldsymbol{u}(t). \tag{4}$$

2.2.2. Selective scan mechanism

Standard SSMs are linear time-invariant (LTI) systems, meaning their parameters remain static regardless of the input. However, accurate physiological grading requires the model to dynamically focus on relevant features. To achieve this, the Selective Scan Mechanism (S6) proposed by Gu and Dao (2024) is employed. This mechanism renders the parameters B, C, and Δ as functions of the input x, allowing the model to selectively propagate or forget information based on the current context. This capability is analogous to the gating mechanisms in LSTMs but operates with significantly higher efficiency (Xie *et al.*, 2025).

2.2.3. 2D Selective scan for spatial features

A fundamental challenge in applying SSMs to computer vision is the non-sequential nature of 2D images. Standard 1D scanning fails to capture the spatial continuity of moss growth patterns. To bridge this gap, the VMamba architecture introduced by Liu *et al.* (2024) is adopted, which utilizes the 2D selective scan (SS2D) module.

SS2D decomposes the input feature map into four distinct scanning paths: top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right. These four sequences are processed in parallel by independent S6 blocks and then merged (Liu *et al.*, 2024). This allows our model to robustly analyze the overall health status of the moss samples without the computational burden associated with global self-attention mechanisms seen in Vision Transformers (Dosovitskiy, 2020). Given the isotropic nature of the moss canopy texture under a 90° top-down perspective, this four-way scanning mechanism effectively eliminates directional bias through multi-path information compensation, enhancing robustness for irregularly growing biological samples.

2.3. Multi-scale feature initialization *via* inception stem

Standard ResNet (Hendrawan *et al.*, 2021) architectures typically initiate feature extraction with a single 7×7 convolution layer followed by max-pooling. While effective for general object recognition, this design assumes that the optimal initial features can be captured at a single fixed spatial scale1. For complex biological samples like moss, where diagnostic features range from microscopic spore textures (Milis *et al.*, 2025) to macroscopic colony shapes, this single-scale initialization may result in the early loss of fine-grained details.

To overcome this limitation, the initial convolution layer of the ResNet50 backbone is replaced with a custom Inception Stem module. As proposed in the Inception-v4 architecture by Szegedy *et al.* (2017), employing a multi-branch stem allows the network to process the input image at different receptive fields simultaneously before entering the deeper residual stages.

As demonstrated in Fig. 2, the inception stem accepts a 3-channel input and decomposes the feature extraction into four parallel pathways, each producing 16 channels:

1. The pooling branch (branch1): A 1×1 convolution followed by a 3×3 MaxPool. This branch focuses on preserving the most dominant structural features (texture intensity) while reducing spatial resolution.
2. The shallow convolution branch (branch2): A sequence of 1×1 and 3×3 convolutions. This path acts as a standard feature extractor for medium-frequency details.
3. The deep convolution branch (branch3): A deeper stack (1×1 → 3×3 → 3×3) with stride 2. By stacking two 3×3 convolutions, this branch effectively approximates a 5×5 receptive field with reduced parameters, capturing larger spatial patterns.
4. The large kernel branch (branch4): A 7×7 convolution. It retains the capability of the original ResNet stem to capture broad, low-frequency global information.

The outputs of these four branches are concatenated along the channel dimension, resulting in a unified feature map of 64 channels. This aggregated output passes through Batch Normalization and ReLU activation, ensuring that the subsequent ResNet layers receive a rich, multi-scale feature representation. This modification allows the model to "see" both fine textures and coarse shapes immediately from the first layer, significantly enhancing the robustness of the subsequent feature encoding.

2.4. Strategic loss function design for imbalanced ordinal data

Standard classification objectives like Cross-Entropy Loss treat all classes as independent and equally distributed entities. However, the physiological grading task presents two distinct statistical challenges that necessitate a more tailored approach. The first challenge is class imbalance, where samples of intermediate health grades significantly outnumber those of extreme grades such as completely necrotic or perfectly healthy tissues (Johnson and Khoshgoftaar, 2019). The second challenge is the ordinal nature of the data, which implies that the health grades possess a natural ordering (Miftahushudur *et al.*, 2025). For instance, misclassifying a Level 1 sample as Level 4 is qualitatively worse than misclassifying it as Level 2. To strictly enforce these constraints during training, a composite loss function is proposed.

2.4.1. Class-weighted re-balancing

To counteract the bias towards majority classes, a class-weighted cross-entropy loss (Cui *et al.*, 2019) is employed. A weight vector based on the inverse class frequency is calculated in the training set. This weight w_i for a specific class i is mathematically defined as:

$$w_i = \frac{N}{CN_i}. \tag{5}$$

In this equation, N represents the total number of samples in the dataset, C denotes the total number of distinct classes, and N_i signifies the number of samples belonging to class i. By modulating the loss contribution of each sample with this calculated weight, the network is penalized more heavily for misclassifying minority classes. These minority classes are typically the critical diseased or perfect samples that define the extremes of the grading scale (Ayoub *et al.*, 2025).

2.4.2. Ordinal regression constraint

To embed this ordinality into the optimization landscape, an ordinal distance penalty is introduced (Beckham and Pal, 2017). The total loss function L_{total} is formulated as a weighted sum of the weighted classification loss and the ordinal distance penalty:

$$L_{total} = \lambda_{cls} L_{WCE}(y, \hat{y}) + \lambda_{ord} L_{dist}(y, \hat{y}), \tag{6}$$

here: L_{WCE} represents the weighted cross-entropy loss and L_{dist} denotes the distance-based penalty term. The hyperparameters λ_{cls} and λ_{ord} are utilized to control the trade-off between classification accuracy and ordinal consistency. In our experimental setup, to prioritize the primary classification objective while providing a sufficient constraint against large-margin ordinal errors, these hyperparameters were empirically set to $\lambda_{cls} = 1.0$ and $\lambda_{ord} = 0.3$. Specifically, since growth grades are defined within Eq. (1, 4), the maximum squared deviation $|y - \hat{y}|^2$ is strictly bounded by 9, which, when scaled by $\lambda_{ord} = 0.3$, constrains the penalty to a maximum of 2.7 to ensure gradient stability; furthermore, the class weights w_i are normalized such that $\sum(wi\, Ni/N) = 1$ to prevent fluctuations in the loss scale.

3. EXPERIMENTS

3.1. Experimental setups

3.1.1. Model architecture and training settings

To ensure rigorous reproducibility and explicitly demonstrate the architectural efficiency of VSS-ORNet, all experiments were conducted on a resource-constrained mobile workstation equipped with an Intel Core i7-12700H CPU, 16GB of system RAM, and an NVIDIA GeForce

RTX 3060 Laptop GPU. The computational environment was standardized on Windows 11 (64-bit) utilizing Python 3.12, PyTorch 2.5.1, and CUDA 12.1. Crucially, the capability to optimize this hybrid framework efficiently on consumer-grade mobile hardware directly substantiates our claims of parameter efficiency. By strategically replacing dense, large-kernel convolutions with a multi-branch Inception Stem and implementing $\mathcal{O}(N)$ linear-complexity VMamba blocks, VSS-ORNet achieves a highly compact computational footprint.

During preprocessing, all images were uniformly standardized to a resolution of 224×224 pixels utilizing OpenCV (4.12.0) and auxiliary libraries (Diffusers 0.36.0, Ultralytics 8.4.2). To artificially expand the limited feature space of our dataset (1 551 images), we strictly applied stochastic geometric regularizations. These techniques included bounded random rotations (-90° to 90°) and orthogonal flipping (p = 0.5) to enforce robust spatial invariance. The architecture was trained end-to-end over 50 epochs using the Adam optimizer. To stabilize gradient flow and prevent rapid memorization of background noise, we enforced a highly conservative, fixed learning rate of 1×10^{-5}. This constrained optimization trajectory is dynamically governed by our customized weighted ordinal loss, which inherently neutralizes class imbalance while penalizing rank-inconsistent predictions to preserve the biological ordinality of the physiological grading. The dataset was partitioned prior to applying an online dynamic augmentation pipeline that generated over 54 000 stochastic geometric variations across 50 epochs, synergizing with the multi-scale Inception Stem and strict channel-wise normalization to cultivate robust invariant representations against environmental lighting, scale, and viewpoint noise.

3.1.2. Evaluation

The grading performance is evaluated on the held-out test set using a comprehensive suite of metrics. Top-1 Accuracy is employed for general correctness, while Macro and Weighted F1-Scores are used to address class imbalance. For ordinal consistency, reflecting the natural progression of moss health from Level 1 to 4, quadratic weighted Kappa (QWK) and mean absolute error (MAE) are utilized. Additionally, the computational efficiency profile, including inference speed (FPS), latency (ms), FLOPs, and total parameter count, evaluated with a batch size of 16 on an NVIDIA GeForce RTX 3060 Laptop GPU, as detailed in Table 2, is reported to assess real-time deployment feasibility.

3.2. Main results

3.2.1. Ablation studies

To rigorously validate the contribution of each proposed component and eliminate the unreliability of a single train/test split, a step-by-step ablation study using a ResNet50 baseline is conducted under a strict 5-fold cross-validation scheme. The quantitative performance metrics, reported as mean ± standard deviation, are detailed in Table 3. Complementing the static performance indicators, the training dynamics illustrated in Fig. 3 provide further insight into the optimization efficiency of the proposed VSS-ORNet. As observed in Fig. 3e), our model exhibits a significantly accelerated convergence rate compared to the Baseline a) and other intermediate variants. Specifically, the accuracy curve of the proposed model ascends steeply within the first 10 epochs and rapidly plateaus at a high precision level, whereas other models require more epochs to stabilize.

As shown in Table 3, the Baseline model achieves an accuracy of 91.63±0.34% and a QWK of 0.9409±0.0013. Improvements were introduced incrementally to observe their individual and collective impacts. Because Accuracy can be misleading for ordinal problems, our analysis prioritizes QWK and MAE. (i) Replacing standard Cross-Entropy with the Weighted Ordinal Loss significantly reduces the MAE from 0.1009 to 0.0837 and elevates the QWK to 0.9502±0.0028. This substantiates that explicitly penalizing ordinal deviations effectively regularizes the optimization landscape, mitigating the large-margin misclassifications inherent in scalar grading tasks. (ii) The integration of the inception module lifts accuracy to 92.49±0.26%. This confirms that parallel multi-kernel branches are essential for resolving scale-variant features that single-scale

Table 2. Experimental environment specifications

Parameter category	Configuration
Operating system	Windows 11 (64-bit)
Processing unit (CPU)	Intel Core i7-12700H
Graphics accelerator (GPU)	NVIDIA GeForce RTX 3060 Laptop
System memory (RAM)	16GB
Deep learning stack	PyTorch 2.5.1 CUDA 12.1
Language and libraries	Python 3.12, Diffusers 0.36.0, OpenCV 4.12.0, Ultralytics 8.4.2
Input resolution	224×224 pixels

Table 3. Quantitative ablation study

Model	Ordinal loss	Inception	VMamba	Accuracy (%)	Macro-F1	QWK	MAE
Baseline	F	F	F	91.63 ± 0.34	0.9099 ± 0.0025	0.9409 ± 0.0013	0.1009 ± 0.0034
Model1	T	F	F	92.15 ± 0.23	0.9115 ± 0.0056	0.9502 ± 0.0028	0.0837 ± 0.0027
Model2	T	T	F	92.49 ± 0.26	0.9172 ± 0.0032	0.9463 ± 0.0030	0.0901 ± 0.0029
Model3	T	F	T	92.36 ± 0.45	0.9127 ± 0.0041	0.9563 ± 0.0035	0.0858 ± 0.0045
Proposed model	T	T	T	**93.49 ± 0.54**	**0.9266 ± 0.0075**	**0.9662 ± 0.0042**	**0.0830 ± 0.0033**

Metrics include accuracy, Macro-F1 Score, quadratic weighted Kappa, and mean absolute error. “T” and “F” denote the presence and absence of a component.

convolutions overlook. (iii) Separately integrating the VMamba Block yields a superior QWK of 0.9563±0.0035. This performance gain validates the 2D Selective Scan mechanism’s ability to model long-range spatial dependencies, effectively perceiving the morphological continuity of the moss bed. (iv) Combining all modules yields the best and most stable performance across all metrics, achieving 93.49±0.54% accuracy and an optimal 0.9662±0.0042 QWK. The MAE drops to the lowest level of 0.0830±0.0033, demonstrating that the synergy between local multi-scale extraction and global state-space modeling provides the most robust representation for physiological grading.

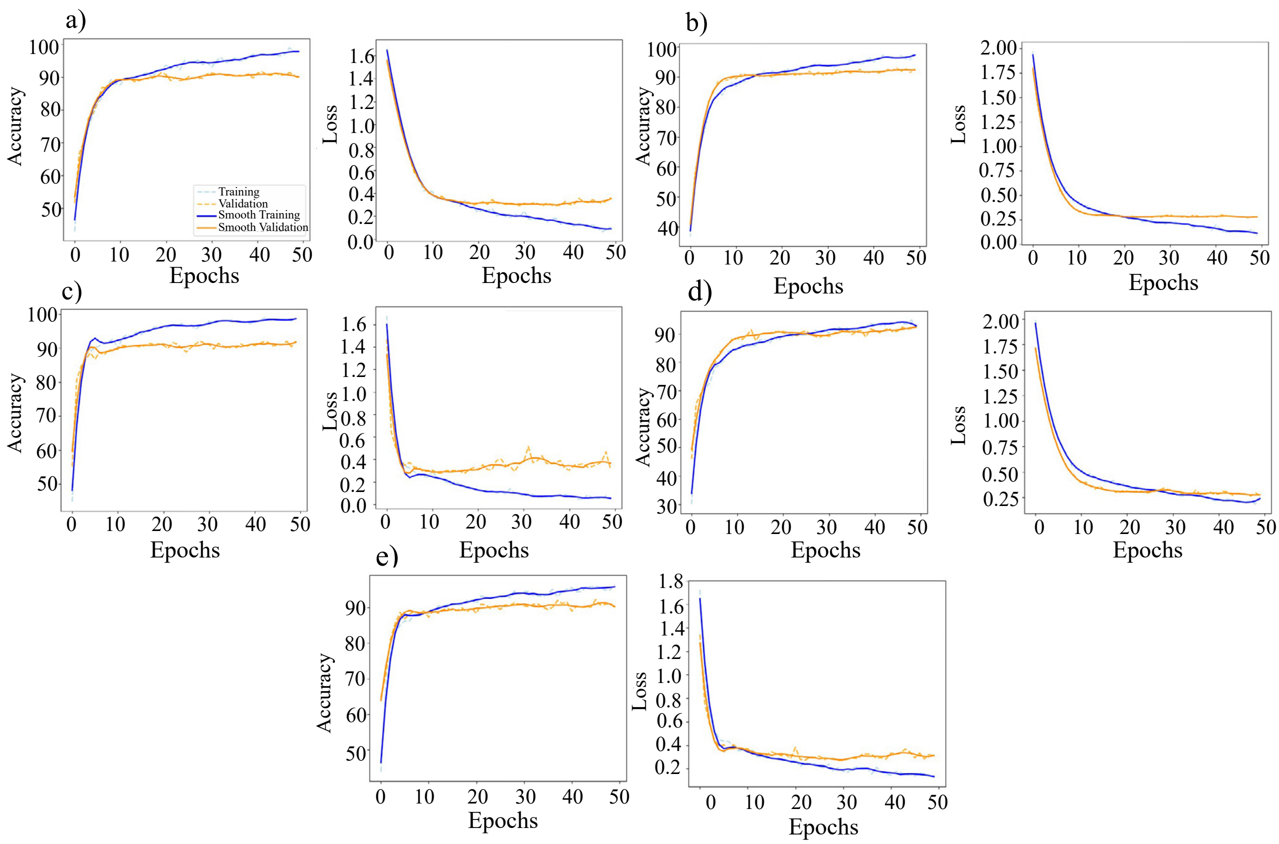


Fig. 3. Training and validation performance curves. The sub-figures correspond to the stepwise integration of components: a) baseline; b) Model 1 (baseline + ordinal oss); c) Model 2 (baseline + ordinal loss + inception stem); d) Model 3 (baseline + ordinal loss + VMamba); e) proposed VSS-ORNet. The solid lines represent smoothed curves, while the dashed lines indicate raw data.

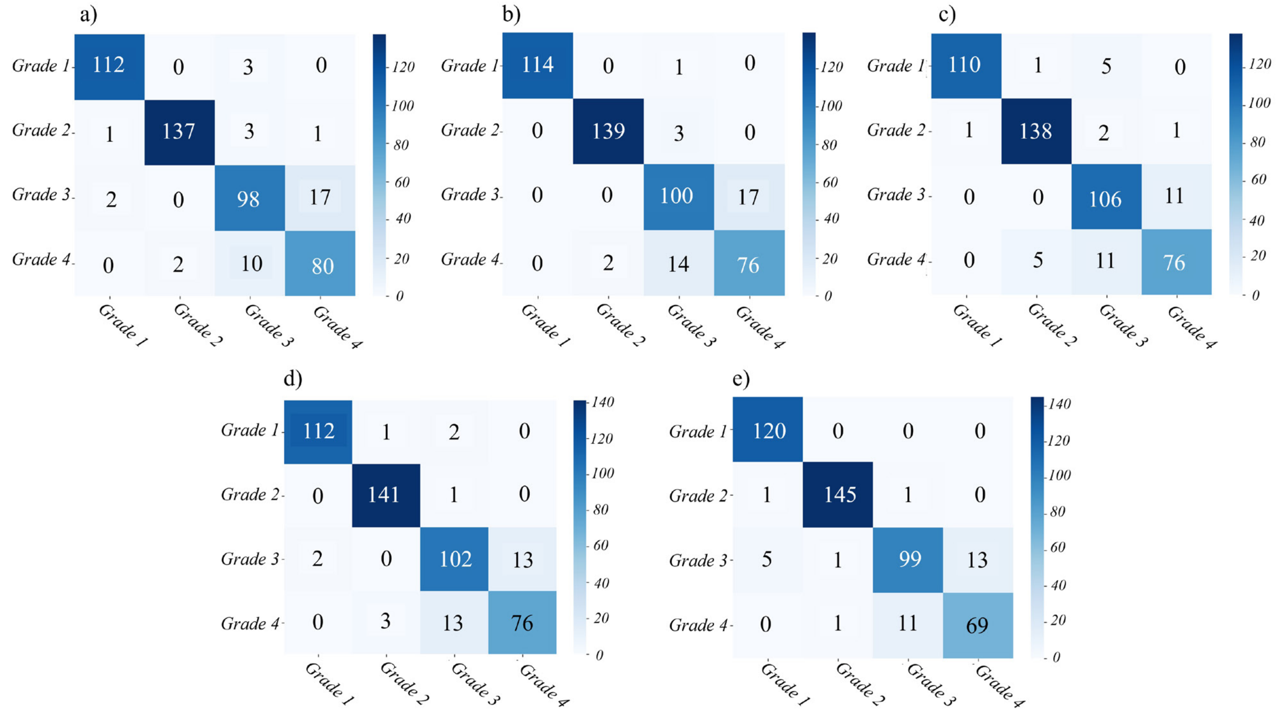


Fig. 4. Confusion matrices: a) baseline; b) model 1 (baseline + ordinal loss); c) model 2 (baseline + ordinal loss + inception stem); d) model 3 (baseline + ordinal loss + VMamba); e) proposed VSS-ORNet.

The confusion matrices of the best-performing fold in Fig. 4 provide a granular view of the model's discriminative capability across the 466 test samples. As observed, the early variants exhibit noticeable off-diagonal noise, indicating instances of severe large-margin misclassification. Specifically, in Fig. 4c Model 2, the network struggled with inter-class ambiguity, misclassifying 5 instances of Grade 4 as Grade 2 and 4 instances of Grade 1 as Grade 3. These represent critical diagnostic failures spanning multiple growth phases. However, with the sequential integration of the proposed modules, these specific errors are systematically minimized. Crucially, in Fig. 4e the Proposed VSS-ORNet, severe cross-grade errors are substantially constrained. This visual evidence reinforces that the Weighted Ordinal Loss effectively constrains predictions to the diagonal vicinity. While occasional misclassifications remain, they are almost exclusively restricted to the adjacent growth stage, which is biophysically acceptable given the visually ambiguous transitional states of moss. The global context modeling of Mamba resolves this ambiguity far better than baselines, ensuring robust and consistent grading.

To explicitly verify whether the network captures the true biophysical characteristics of moss degradation rather than spurious visual artifacts, Grad-CAM heatmaps (Fig. 5) are employed to analyze the model's spatial focus. As observed, standard CNN baselines (Fig. 5b, c) exhibit heavily localized Effective Receptive Fields, fixating almost exclusively on isolated morphological anomalies. This localized bias is biologically misaligned with the nature of *Hypnum* moss, where physiological stress-such as progressive desiccation or chlorosis-manifests as continuous, dispersed phytosanitary shifts across the entire colony.

The proposed VSS-ORNet (Fig. 5f) generates a uniformly distributed, global ERF that comprehensively covers the entire moss canopy, including peripheral structures. More importantly, validating the model's interpretability, the high-activation "red zones" in the VSS-ORNet heatmaps precisely correlate with actual physical symptoms of moss degradation. Instead of fixating on random background noise, the model correctly focuses its highest attention on the progressing edges of localized canopy shrinkage and the dispersed early-stage browning of gametophytes. By mathematically integrating these non-local macroscopic variations instead of being deceived by localized focal points, the model aligns with the requirements of bulk quality control. Practically, this global spatial awareness ensures that automated inspection systems in dynamic storage facilities and high-speed production lines can reliably assess the holistic health of a plant tray. This biophysically-

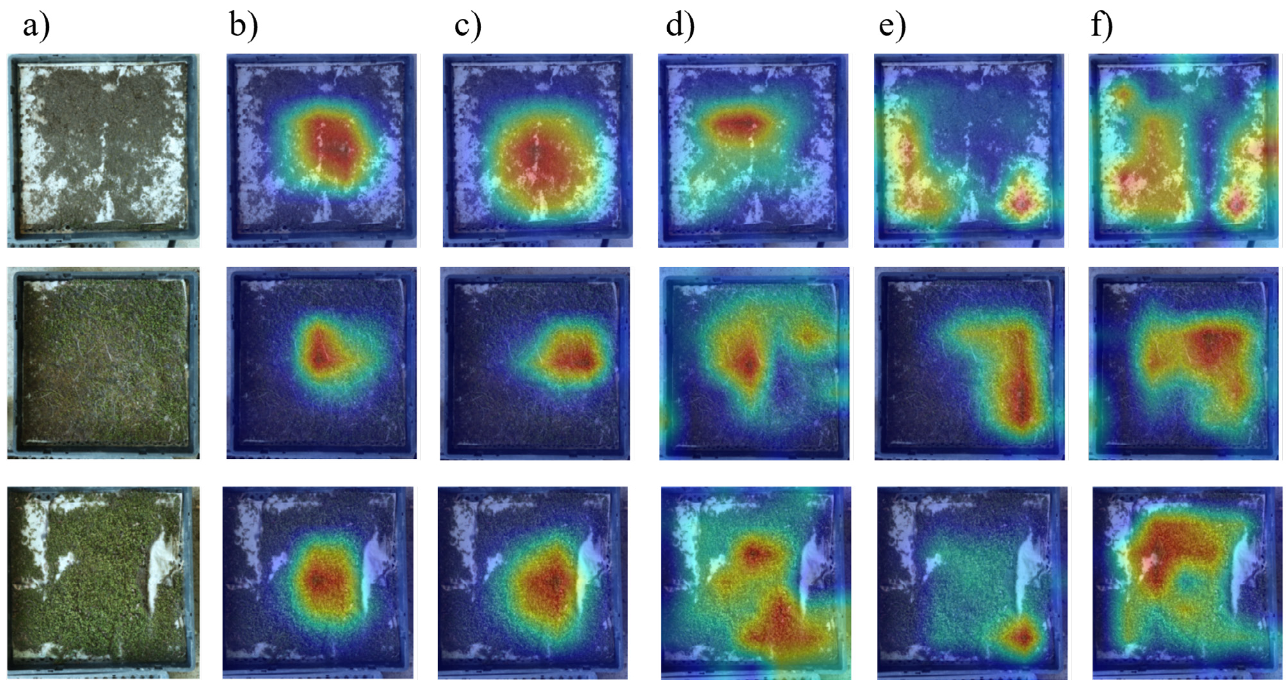


Fig. 5. Grad-CAM visualization of effective receptive fields across model variants. Red regions denote high activation intensity: a) original image; b) baseline; c) model 1 (baseline + ordinal loss); d) model 2 (baseline + ordinal loss + inception stem); e) model 3 (baseline + ordinal loss + vmamba); f) proposed VSS-ORNet.

grounded global coverage is the fundamental reason behind the model's success in eradicating large-margin diagnostic errors during automated grading.

3.2.2. Comparative analysis with state-of-the-art models

To comprehensively evaluate the effectiveness and unique contributions of our proposed VSS-ORNet, we benchmark it against five mainstream and widely adopted deep learning architectures, each representing a distinct paradigm in computer vision. The expanded comparative baselines span both convolutional and transformer paradigms, including VGG16 (Simonyan and Zisserman, 2014) representing standard deep sequential convolutions, DenseNet121 (Huang *et al.*, 2017) for dense feature reuse, Inception-ResNet-v2 (Szegedy *et al.*, 2017) for multi-scale residual integration, EfficientNet-B4 (Tan and Le, 2019) for optimized compound scaling, ConvNeXt (Liu *et al.*, 2022) representing modernized pure convolutional macro-designs, alongside ViT-Base (Dosovitskiy, 2020) representing global patch-based self-attention, and Swin-T (Liu *et al.*, 2021) for hierarchical shifted-window transformer architectures. Table 3 presents a detailed comparison of these models across classification metrics, including Accuracy, Macro-F1, Quadratic Weighted Kappa, and Mean Absolute Error, as well as computational efficiency metrics such as Parameters, FLOPs, and Frames Per Second.

As detailed in Table 4, VSS-ORNet reliably achieves state-of-the-art performance across all cross-validation metrics, reaching 93.49% accuracy, a 0.9662 QWK, and a minimized mean absolute error of 0.0830. This comfortably outperforms both top-tier CNNs (e.g., inception-ResNet-v2 at 91.24%) and newly benchmarked vision transformers, including Swin-T (90.85%) and ViT-Base (88.62%, MAE 0.1085). These comparative results expose the inherent limitations of standard architectures under data-scarce conditions: massive CNNs (e.g., VGG16) fail to capture dispersed, non-local desiccation patterns across the moss canopy, while global self-attention mechanisms in pure transformers prove overly "data-hungry," leading to sub-optimal feature learning on our limited dataset of 1 551 images.

From a practical deployment perspective, VSS-ORNet strikes a highly favourable computational balance for agricultural edge devices. Requiring 49.90 million parameters and 5.53 G FLOPs, it avoids the extreme memory bloat of standard global Transformers like ViT-Base (86.56 M parameters, 17.58 G FLOPs). Furthermore, while lightweight alternatives like Swin-T and pure convolutional models achieve inference speeds exceeding 68 and 100 frames per second respectively, they do so at an unacceptable cost to ordinal grading reliability. For real-world automated inspection lines-where the priority is reliable, batch-wise health assessment rather than ultra-high-speed video streaming-our model's processing rate of 29.36

Table 4. Quantitative benchmarking of the proposed VSS-ORNet against state-of-the-art deep learning architectures

Model	Accuracy (%)	Macro-F1	QWK	MAE	Parameters	FLOPs	FPS(f/s)
Inception-ResNet-v2	91.24 ± 0.26	0.8925 ± 0.0138	0.9325 ± 0.0086	0.0956 ± 0.0042	54.25 M	6.47 G	16.27
ConvNeXt	88.75 ± 0.72	0.8264 ± 0.0096	0.8916 ± 0.0134	0.1056 ± 0.0079	27.81 M	4.46 G	**106.61**
EfficientNet-B4	85.53 ± 0.61	0.9015 ± 0.0081	0.8865 ± 0.0067	0.0925 ± 0.0034	17.56 M	**1.58 G**	30.52
VGG16	90.26 ± 0.85	0.9123 ± 0.0064	0.9359 ± 0.0051	0.0967 ± 0.0031	134.29 M	15.52 G	106.14
DenseNet121	86.37 ± 0.23	0.8968 ± 0.0049	0.8731 ± 0.0063	0.1102 ± 0.0059	**6.96 M**	2.90 G	35.93
VSS-ORNet	**93.49 ± 0.54**	**0.9266 ± 0.0075**	**0.9662 ± 0.0042**	**0.0830 ± 0.0033**	49.90 M	5.53G	29.36

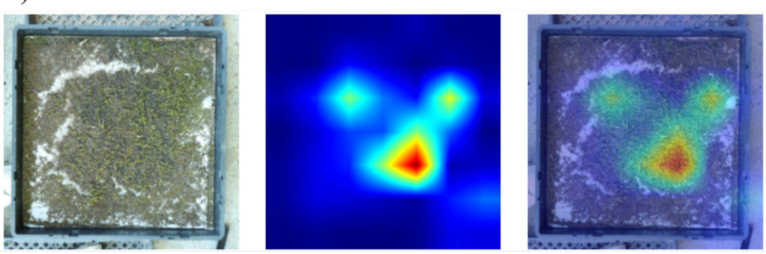
frames per second is more than sufficient. Ultimately, this confirms that VSS-ORNet successfully translates advanced visual state-space modeling into a highly viable, economically beneficial tool for precision bryophyte cultivation.

3.2.3. Failure case analysis

Despite the high overall accuracy and robust ordinal consistency achieved by VSS-ORNet, analyzing its residual errors provides valuable insights into its current boundary limitations. Figure 6 presents typical failure cases encountered during testing alongside their Grad-CAM visualizations.

The primary causes for these specific errors can be categorized into two main biophysical and visual challenges. First, moss exhibits a highly continuous biological degradation process, frequently resulting in mixed transitional states (Fig. 6a). In this instance, a Grade 3 sample is misclassified as Grade 2. The Grad-CAM overlay reveals that the model's attention is strongly drawn to a concentrated patch of residual green gametophytes, while ignoring the surrounding browning areas. Because healthy and stressed tissues are densely interspersed, the macroscopic representation averages out the fine-grained diagnostic features, leading to a conservatively optimistic prediction. Second, uneven illumination and localized shadows (Fig. 6b) can occasionally obscure the true pigmentation of the canopy. In this case, a dormant Grade 4 sample is mispredicted as Grade 3. The varying light intensity and shadows artificially alter the perceived texture and colour profile, causing the network to misinterpret the desiccated state. Recognizing these specific visual ambiguities underscores the inherent limitations of reliance on purely RGB sensors and highlights the necessity for future systems to integrate multi-angle lighting compensation or multi-modal sensors.

a)



b)

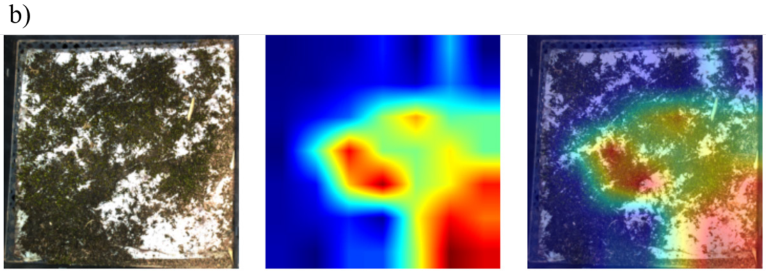


Fig. 6. Typical failure cases of VSS-ORNet with Grad-CAM visualizations. a) Ground truth: Grade 3, predicted: Grade 2; b) Ground truth: Grade 4, predicted: Grade 3.

4. CONCLUSIONS

The critical bottleneck of automated quality control in *Hypnum* moss bio-production was addressed in this study. Moving beyond the pure methodological optimization of computer vision, the findings provide profound biophysical and practical implications for precision agriculture. Biologically, the physiological degradation of *Hypnum* moss is not a set of discrete, isolated events, but a continuous morphological spectrum. Standard single-stream CNN classifiers fail to capture this biophysical reality, resulting in unacceptably high error margins, as evidenced by DenseNet121 recording a Mean Absolute Error (MAE) of 0.1102. By introducing the Weighted Ordinal Loss, our VSS-ORNet explicitly aligns the optimization landscape with the actual phytosanitary degradation curve. This ordinal regression strictly penalizes large-margin diagnostic errors, successfully reducing the MAE to a highly robust 0.0830. This biologically ensures that misclassifications are safely confined to adjacent, visually ambiguous growth stages, preventing disastrous grading mistakes in automated storage systems.

Furthermore, the integration of the Visual State Space block provides a highly interpretable mechanism for understanding moss physiology. Unlike standard convolutions that may overfit to localized surface anomalies, the global receptive field of our architecture effectively captures non-local morphological symptoms, such as widespread canopy shrinkage or distributed pigment degradation across the colony. Economically, this demonstrates that spatially aware, low-cost RGB imaging can serve as a highly viable alternative to capital-intensive hyperspectral systems. Achieving a reliable, cross-validated top-1 accuracy of 93.49%, this RGB-based approach significantly lowers the capital expenditure for automated inspection systems, facilitating scalable, non-invasive quality control for large-scale bryophyte cultivation facilities.

Despite these advancements, the real-world applicability of this study must be contextualized within its inherent limitations. First, the current model is only validated under controlled indoor conditions where environmental variables are strictly regulated. Consequently, its generalization to complex field environments needs improvement, as dynamic factors like fluctuating illumination, canopy shadows, and varying surface humidity in open-field or semi-open greenhouse conditions can alter moss's reflectance and exacerbate visual ambiguity. Second, the system cannot detect internal water stress from RGB alone. A targeted failure case analysis reveals that misclassifications predominantly occur during the pre-symptomatic phase; because standard RGB cameras only capture surface-level morphological changes, deep-tissue physiological stress often goes undetected until macroscopic visual symptoms manifest (Wakamori *et al.*, 2020). Finally, regarding industrial deployment, the model's inference speed of 29.36 FPS-while adequate for standard batch inspection and storage monitoring-lags significantly behind the 106.61 FPS throughput of lightweight macro-architectures like ConvNeXt (Liu *et al.*, 2022), presenting a potential computational bottleneck for ultra-high-speed conveyor sorting lines.

To transition VSS-ORNet from theoretical validation to robust deployment in high-speed agricultural inspection lines and dynamic storage facilities, future research must address the method's current limitations under realistic "field" conditions. First, systematic stress testing is imperative to guarantee diagnostic resilience against physical environmental noise, such as severe greenhouse illumination fluctuations and motion blur inherent to automated sorting conveyors. Second, current descriptive visual diagnostics must advance to quantitative interpretability. This will mathematically validate that the network consistently isolates true phytosanitary degradation-such as non-local desiccation patterns or macroscopic canopy shrinkage-rather than spurious background artifacts. Finally, because surface-level RGB imaging inherently fails to capture pre-symptomatic internal water stress, future agricultural edge-computing systems should prioritize low-cost multi-modal sensor fusion. Integrating thermal imaging to detect early-stage transpirational cooling anomalies (Fevgas *et al.*, 2025) will overcome these biophysical limitations. Furthermore, the proposed 2D diagnostic framework can be synergized with advanced 3D vision technologies for automated robotic deployment. For instance, geometry-aware 3D point cloud learning (Wang *et al.*, 2025) and specialized structural recognition robots (Hu *et al.*, 2024) provide potential pathways for achieving precise multi-dimensional health monitoring in unstructured agricultural or industrial environments.

Authors' contributions: Research concept and design: Z.L., A.F., and M.C.; collection and/or assembly of data: Z.L.; data analysis and interpretation: Z.L. and A.F.; writing the article: Z.L. and A.F.; critical revision of the article: Z.L. All authors have read and agreed to the published version of the manuscript.

Conflict of interest. The authors declare no conflicts of interest.

5. REFERENCES

Anami, B.S., Malvade, N.N., Palaiah, S., 2020. Deep learning approach for recognition and classification of yield affecting paddy crop stresses using field images. Artificial Intelligence Agric. 4, 12-20. https://doi.org/10.1016/j.aiia.2020.03.001

Anderson, L.E., Crum, H.A., Buck, W.R., 1990. List of the mosses of North America north of Mexico. Bryologist 448-499. https://doi.org/10.2307/3243611

Ayoub, S., Baig, I., Ashraf, M., Okasha, M., 2025. Impact of dataset quality on deep learning models for dragon fruit and leaf health classification. Impact Agric. 1, https://doi.org/10.65500/agriculture-2025-001

Beckham, C., Pal, C., 2017. Unimodal probability distributions for deep ordinal classification. Int. Conf. Machine Learning 411-419. https://doi.org/10.48550/arXiv.1705.05278

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., Belongie, S., 2019. Class-balanced loss based on effective number of samples. Proc. IEEE/CVF Conf. Computer Vision Pattern Recognition 9268-9277. https://doi.org/10.1109/CVPR.2019.00949

Dosovitskiy, A., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

Fevgas, G., Lagkas, Papadopoulos, T., Sarigiannidis, P., Argyriou, V., 2025. Integrating thermal infrared and RGB imaging for early detection of water stress in lettuces with comparative analysis of IoT sensors. Smart Agric. Technol. 11, 100881. https://doi.org/10.1016/j.atech.2025.100881

Furbank, R.T., Tester, M., 2011. Phenomics-technologies to relieve the phenotyping bottleneck. Trends Plant Sci. 16, 635-644. https://doi.org/10.1016/j.tplants.2011.09.005

Glime, J., 2017. Bryophyte Ecology. Michigan Technological University, Michigan.

Gu, A., Dao, T., 2024. Mamba: Linear-time sequence modeling with selective state spaces. Ist Conf. Language Modeling, https://doi.org/10.48550/arXiv.2312.00752

Gu, A., Goel, K., Ré, C., 2021. Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396. https://doi.org/10.48550/arXiv.2111.00396

Gulzar, Y., 2023. Fruit image classification model based on MobileNetV2 with deep transfer learning technique. Sustainability 15, 1906. https://doi.org/10.3390/su15031906

Gulzar, Y., 2025a. Applications of transfer learning in sunflower disease detection: advances, challenges, and future directions. Turk. J. Biol. 49, 534-549. https://doi.org/10.55730/1300-0152.2763

Gulzar, Y., 2025b. Papaya leaf disease classification using pretrained deep learning models: A comparative study. Appl. Fruit Sci. 67, 287. https://doi.org/10.1007/s10341-025-01533-1

Gulzar, Y., 2025c. PapNet: An AI-driven approach for early detection and classification of Papaya leaf diseases. Appl. Fruit Sci. 67, 256. https://doi.org/10.1007/s10341-025-01466-9

Gulzar, Y., Ünal Z., Şahbaz, K., Alkanan, M., 2025. PTL-Inception: integrating deep learning and taxonomy for desert plant classification. Diversity 17, 806. https://doi.org/10.3390/d17110806

Gündüz, H., Günal, S., 2024. A lightweight convolutional neural network (CNN) model for diatom classification: DiatomNet. PeerJ Comput. Sci. 10, e1970. https://doi.org/10.7717/peerj-cs.1970

Haque, Z. Koklu, M., Mirza, M.A., Omar, M., Mukhayyokhon, S., 2025. Mitigating Spatial Scale Loss in CNN-Based Fine-Grained Image Classification: Application to Date Fruit Grading. Impact in Agriculture 1. https://doi.org/10.65500/agriculture-2025-005

Hendrawan, Y., Damayanti, R., Firmanda Al Riza, D., Bagus Hermanto, M., 2021. Classification of water stress in cultured Sunagoke moss using deep learning. Telkomnika (Telecommunication Computing Electronics and Control) 19, 1594-1604. https://doi.org/10.12928/telkomnika.v19i5.20063

Hu, K., Chen, Z., Kang, H., Tang, Y., 2024. 3D vision technologies for a self-developed structural external crack damage recognition robot. Automation in Construction 159, 105262. https://doi.org/10.1016/j.autcon.2023.105262

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. Proc. IEEE Conf. Computer Vision And Pattern Recognition 4700-4708. https://doi.org/10.1109/CVPR.2017.243

Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J. Big Data 6, 1-54. https://doi.org/10.1186/s40537-019-0192-5

Kamilaris, A., Prenafeta-Boldú, F.X., 2018. Deep learning in agriculture: A survey. Computers Electronics Agriculture 147, 70-90. https://doi.org/10.1016/j.compag.2018.02.016

Liakos, K.G., Busato, P., Moshou, D., Pearson, S., Bochtis, D., 2018. Machine learning in agriculture: A Review. Sensors (Basel) 18, 2674. https://doi.org/10.3390/s18082674

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., *et al*., 2024. Vmamba: Visual state space model. Advances In Neural Inf. Proces. Systems 37, 103031-103063. https://doi.org/10.52202/079017-3273

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., *et al*., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. Proc. IEEE/CVF Int. Conf. Computer Vision 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986

Liu, Z., Mao, H., Wu, C-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 11976-11986. https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html

Miftahushudur, T., Miftahushudur, T., Mertkan Sahin, H., Grieve, B., Hujun Yin, H., 2025. A survey of methods for addressing imbalance data problems in agriculture applications. Remote Sensing 17, 454. https://doi.org/10.3390/rs17030454

Milis, A., Hofmann, M., Mäder, P., Wäldchen, J., de Haan, M., Ballings, P., 2025. Towards the automatized identification of moss species from their spore morphology. Annals of Botany mcaf215. https://doi.org/10.1093/aob/mcaf215

Oliver, M.J., Velten, J., Mishler, B.D., 2005. Desiccation tolerance in bryophytes: a reflection of the primitive strategy for plant survival in dehydrating habitats? Integr. Comp. Biol. 45, 788-799. https://doi.org/10.1093/icb/45.5.788

Picon, A., Alvarez-Gila, A., Seitz, M., Ortiz-Barredo, A., Echazarra, J., Johannes, A., 2019. Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild. Computers Electronics Agric. 161, 280-290. https://doi.org/10.1016/j.compag.2018.04.002

Proctor, M.C., Oliver, M.J., Wood, A.J., Alpert, P., Stark, L.R., Cleavitt, N.L., *et al*., 2007. Desiccation-tolerance in bryophytes: a review. The Bryologist 110, 595-621. https://doi.org/10.1639/0007-2745(2007)110[595:DIBAR]2.0.CO;2

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. Proc. AAAI Conf. Artificial Intelligence https://doi.org/10.1609/aaai.v31i1.11231

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. Int. Conf. Machine Learning 6105-6114. https://doi.org/10.48550/arXiv.1905.11946

Tustin, A., 1947. A method of analysing the behaviour of linear systems in terms of time series. J. Institution of Electrical Engineers-Part IIA: Automatic Regulators and Servo Mechanisms 94, 130-142. https://doi.org/10.1049/ji-2a.1947.0020

Wakamori, K., Mizuno, R., Nakanishi, G., Mineno H., 2020. Multimodal neural network with clustering-based drop for estimating plant water stress. Computers Electronics Agriculture 168, 105118. https://doi.org/10.1016/j.compag.2019.105118

Walter, A., Liebisch, F., Hund, A., 2015. Plant phenotyping: from bean weighing to image analysis. Plant Methods 11, 14. https://doi.org/10.1186/s13007-015-0056-8

Wang, H., Zhang, G., Cao, H., Hu, K., Wang, Q., Deng, Y., 2025. Geometry-aware 3D point cloud learning for precise cutting-point detection in unstructured field environments. J. Field Robotics 42, 3063-3076. https://doi.org/10.1002/rob.22567

Xie, F., Nie, J., Tang, Y., Zhang, W., Zhao, H., 2025. Mamba-adaptor: State space model adaptor for visual recognition. In: Proc. Conf. Computer Vision Pattern Recognition Conf. 20124-20134. https://openaccess.thecvf.com/content/CVPR2025/html/Xie_Mamba-Adaptor_State_Space_Model_Adaptor_for_Visual_Recognition_CVPR_2025_paper.html

Zangana, H.M., Li, S., Wani, S., 2025. Diffusion models for agricultural imaging: A Systematic Review of Methods, Applications and Future Prospects. Impact in Agriculture 1. https://doi.org/10.65500/agriculture-2025-003