SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

A Hybrid Deep Learning Framework For Multimodal Rice Growth Stage Classification Using Image And Iot Sensor Fusion

Riki Ruli A. Siregar[1] , Kudang Boro Seminar[2] , Sri Wahjuni[3] , Edi Santosa[4]

[1]School of Data Science, Mathematics, and Informatics IPB University, Bogor, Indonesia, Faculty of Energy Telematics Institut Teknologi PLN Jakarta, Indonesia rulirikisiregar@apps.ipb.ac.id
[2]Department of Engineering and Biosystems IPB University Bogor, Indonesia kseminar@apps.ipb.ac.id
[3]School of Data Science, Mathematics, and Informatics IPB University, Bogor, Indonesia my_juni04@apps.ipb.ac.id
[4]Department of Agronomy and Horticulture, Faculty of Agriculture IPB University, Bogor, Indonesia edisang@gmail.com

Abstract: Vertical farming based on IoT is a promising solution to increase rice production in limited urban spaces, but reliable and automated monitoring of crop phenology remains an open challenge. Previous studies have generally relied on unimodal data or simulated environments, which limit accuracy and discrimination power, especially in growth phases with similar visual characteristics. This research aims to develop a multimodal deep learning framework for precise classification of rice growth phases in actual IoT-based vertical farming systems. The proposed method integrates RGB canopy images with temporal environmental sensor data, including temperature, humidity, light intensity, soil moisture, and pH, through end-to-end spatio-temporal feature fusion using several CNN architectures (MobileNet, ResNet-50, VGG-19, and Xception) combined with an LSTM branch before the classification stage. Evaluation was conducted on a real-world dataset annotated by experts and spanning eight rice growth phases, measured in days after planting. Experimental results show that the proposed model significantly outperforms unimodal and CNN-only approaches, achieving a macro average F1 score of 0.96 on the VGG19–LSTM variant and maintaining performance above 0.89 in the most challenging intermediate growth phases, where visual information alone is insufficient. The contribution of these findings to the lightweight MobileNet–LSTM model maintains high accuracy with real-time inference support, making it potentially effective for application in edge computing devices in operational vertical farming systems.

Keywords: Multimodal data fusion, Spatio-temporal analysis, CNN-LSTM, Rice phenology classification, IoT-based vertical farming

I. INTRODUCTION

This study examines the strategic potential of advanced analytical methods for multimodal data, using hybrid deep learning algorithms, in increasing rice production in the proposed application in an IoT-based vertical farming system. This approach aims to address pressing challenges in global food security, rapid

urbanization, and limited agricultural land by improving crop monitoring accuracy and optimizing resource use [1]. The application of deep learning integrated with Internet of Things (IoT) infrastructure has developed as a transformative approach in precision agriculture, increasing crop productivity and resource utilization efficiency, especially in controlled environments such as vertical farming [2]. Rice is one of the world's most important staple food commodities and is traditionally cultivated on large open fields [3], [4]. Traditional rice farming practices make rice crops in paddy fields highly vulnerable to damage from bad weather, pests, and inappropriate irrigation and fertilization methods [5], [6], [7]. The proposed vertical farming method is a good option for cities and other areas with limited space. This method is more efficient in water use, requires less fertilizer, and facilitates year-round planting [8], [9], [10].

One of the biggest challenges in rice cultivation on vertical land is identifying growth stages [11], [12]. Accuracy in this identification is very important for optimal growth phases, irrigation planning, nutrient management (fertilizer based on soil testing) [12], [13], and harvest logistics [14], [15], [16]. In automated agricultural environments, manual observation is generally inefficient and limited. To address this issue, the hybrid deep learning framework in this study combines visual data collection (RGB images) with IoT (Internet of Things) sensor data in real time. This system enables continuous monitoring and adaptive adjustments to the surrounding environment, thereby significantly improving resource distribution efficiency [12]. The hybrid deep learning model's predictive ability can also help identify growth stages, detect growth problems or diseases, and determine the best times to harvest. That will result in increased production and the expansion of sustainable vertical farming [12]. When implemented well, vertical farming will lead to increased yields and sustainable growth.

The integration of computer vision with a sensor-based analysis system has enabled real-time insights into crop performance and growth dynamics, aiding in making timely decisions in predictive action [14]. Recent advances in deep learning, particularly convolutional neural networks (CNNs), have greatly improved crop monitoring and yield prediction. By learning complex representations from large-scale image data--such as those above [15], [16]. Convolutional neural networks are able because they are trained on both synthetic use cases (for which one can find training data) plus actual examples comprising 100 million pixels removed in real inhabitable ecosystems with a wide variety of weather conditions. However, most existing studies are still limited to single inputs either visual images or sensor time series and rely on standalone architectures such as CNNs or LSTMs, which limit their ability to model spatio-temporal dynamics simultaneously [17], [18], [19], [20], [21], [22]. More importantly, end-to-end hybrid deep learning frameworks that explicitly combine image and multimodal sensor data for detailed rice growth analysis in real-world IoT-based vertical farming systems remain largely unexplored. To address this gap, this study proposes an end-to-end multimodal deep learning framework that explicitly combines visual and sensor data to capture spatial and temporal dynamics simultaneously in detailed rice growth stage classification in an operating IoT-based vertical farming environment [23], [24], [25], [26], [27].

This study addresses these weaknesses by proposing an innovative CNN–LSTM hybrid framework that leverages multimodal data from IoT-connected vertical farming systems to continuously and effectively monitor rice growth stages. This design combines spatial feature extraction from CNN with temporal modeling using LSTM. This will improve the accuracy of growth-stage classification and generate more accurate crop yield predictions, enabling real-time plant phenotyping for precision agricultural resource allocation [28], [29].

1.1 Problem statement

The main research questions this study seeks to answer is how implementing the hybrid CNN-LSTM model can improve the accuracy of rice growth-stage classification while also helping to propose an environmentally friendly vertical farming system. To clarify the research context and highlight the key challenges, the main problem statements are summarized as follows:

1. Global food security faces serious challenges due to urbanization, limited arable land, and climate change, necessitating alternative and more efficient agricultural solutions.
2. Conventional rice cultivation relies on large areas of paddy fields and is vulnerable to weather fluctuations, pests, diseases, irrigation, and ineffective fertilization.

3. Although the vertical rice cultivation model promises the use of IoT and AI technologies, it still requires proper monitoring and efficient resource management to optimize conditions at every stage of growth.
4. Accurate identification of each phase of rice plant growth is very important for scheduling irrigation, fertilization, and harvest planning; however, manual methods are not effective and efficient in precision farming.
5. Previous related research has mostly relied on unimodal data (both images and sensor data) with a single model architecture (CNN or LSTM only), which limits the ability to capture comprehensive spatiotemporal dynamics.
6. The research gap lies in the lack of end-to-end hybrid architecture collaboration that integrates CNN-based image analysis and various derivatives of pre-trained and LSTM-based sensor data processing, which enables multimodal monitoring of rice growth data in vertical farming environments.

1.2 Research contribution

This research has broad applications in the fields of precision agriculture and IoT-supported vertical farming:

1. This research proposes an end-to-end hybrid CNN–LSTM architecture that leverages the spatial features of RGB plant images in RGB format and time-series data from IoT-based environmental sensors to classify rice growth stages. The complementary temporal patterns generated by these two types of sensors in different modes slightly improve performance, while the spatio-temporal control device is also an inevitable solution to prevent identification errors caused by a lack of distinguishable morphological changes. Unlike conventional methods based on a single model, the proposed multimodal fusion scheme has good morphological environmental dynamics, which brings significant improvements to the final phenological representation.
2. The introduced tool can enable Vertical Farming and Hybrid Farming through continuous, nondestructive, automated monitoring of rice growth stages. This method eliminates destructive sampling and manual observation, providing high-quality phenological information that does not in any way harm plant health. And so cultivation can proceed continuously throughout the entire cultivation cycle.
3. Large-scale experiments have shown that across all eight stages of standard rice growth, the most significant improvements occur in the middle stages and visually similar stages, such as tillering or panicle initiation, where single-image-based models generally perform poorly. The multimodal CNN–LSTM model provides significantly better classification results.
4. By combining visual image information with real-time environmental sensor data, the proposed model facilitates agronomic decision-making in accordance with ongoing processes. This includes management practices such as flexible irrigation scheduling, soil nutrient diversification, and early detection of growth anomalies at the onset of their emergence.
5. The proposed approach can be applied to vertical farming systems equipped with IoT technology. This framework represents possibilities for sustainable management, automated phenotyping, and vertical farming.

Most existing studies on crop growth stage classification rely on single-modality inputs, such as image-only or sensor-only data, are predominantly conducted in open-field conditions or simulated greenhouse environments or fail to incorporate temporal environmental sensor dynamics in an end-to-end learning framework. As a result, these approaches often struggle to robustly distinguish visually similar phenological stages under controlled farming conditions.

To overcome these limitations, this study proposes a fundamentally different approach. Based on previous research, this study is among the first to introduce an end-to-end multimodal CNN–LSTM framework for detailed phenological classification of rice in a real-world IoT-based vertical farming system, explicitly integrating continuous environmental sensor time series data with visual canopy information via a late-fusion strategy. Unlike conventional feature stacking methods, the proposed architecture simultaneously models spatio-temporal interactions, enabling accurate discrimination of eight standard rice growth stages and supporting environmentally conscious decision-making in vertical farming applications.

II. Method and Material

To implement this study, a proposed IoT-based vertical rice farming system with artificial intelligence was designed. The overall system design is shown in Fig. 1, which consists of layered or tiered cultivation layers, real-time monitoring, and intelligent decision-making based on the data patterns received.

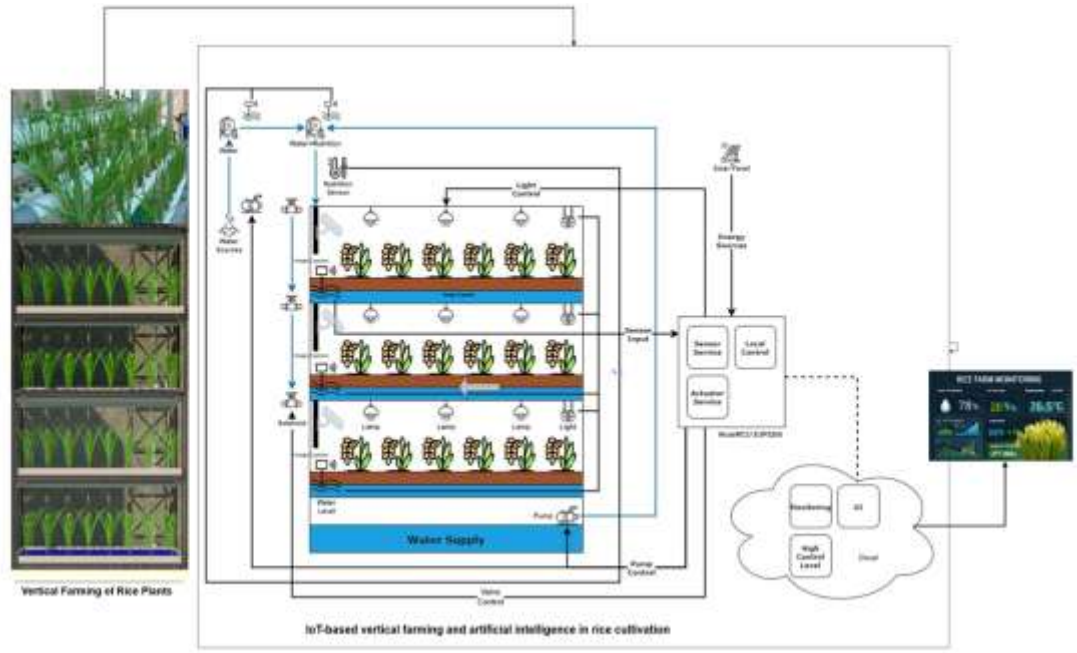


Fig. 1. IoT-based vertical farming architecture integrated with artificial intelligence for precision rice cultivation.

With the proposed system, as shown in Fig. 1, this research will develop an IoT and AI-based vertical farming architecture that supports precision farming in limited land areas, such as urban areas. The layered vertical growth structure is equipped with distributed IoT sensors to record environmental parameters, including temperature, humidity, nutrient concentration, and water level, while image scanning devices continuously monitor photos of rice plants. Sensor and image data are sent to edge controllers UP for local control and forwarded to a cloud-based platform for high-level processing. Deep learning techniques are used to analyze rice growth by combining visual phenological features extracted from plant images with temporal patterns obtained from sensor data. The resulting multimodal insights are used by a decision-support system to automatically regulate irrigation, nutrient delivery, and artificial lighting via a closed-loop control mechanism. The contributions of this work remain threefold in the way that real-time multimodal IoT sensing and deep learning driven decision making are integrated in a vertical rice cultivation system, which allows for fine-grain growth monitoring and adaptive control that goes beyond what has been studied in conventional paddy farming or even existing vertical farming studies.

2.1 Proposed hybrid deep learning framework

This study proposes the use of hybrid deep learning techniques to provide a comprehensive method framework that does not interfere with IoT zoning and, at the same time, integrates all rice plant growth-stage data. Fig. 2 shows that the process flow for this methodology comprises four main stages: Data Collection and Construction, Data Pre-processing, Multi-input Model Development and Training, and Model Evaluation and Implementation. Each stage aims to optimize how information is extracted from features, integrated with other data sources added on top, and how performance in predictive tasks can be improved through real-time monitoring in a controlled vertical farming environment.

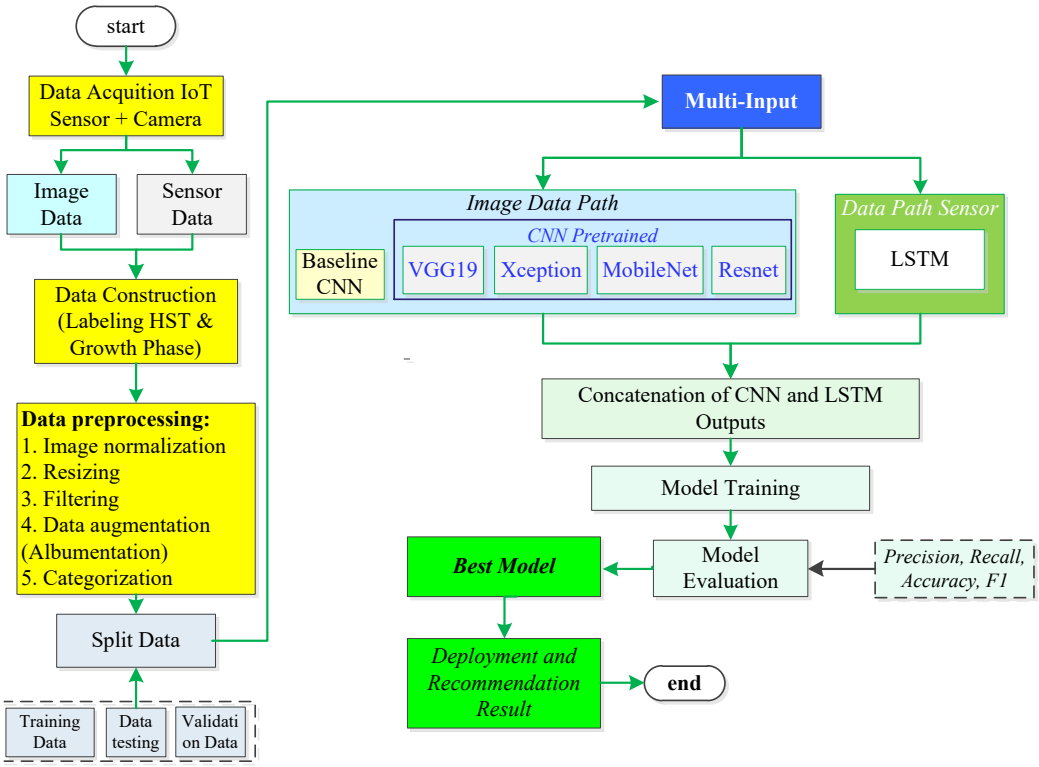


Fig. 2. The workflow of the proposed multimodal deep learning framework for classifying rice growth stages using IoT sensor data and image data.

The hybrid CNN-LSTM algorithm is the focus of the research illustrated in Fig. 2. CNNs are used for image processing, while LSTMs analyze sensor-based time series data via a concatenation-based late-fusion mechanism. Thus, this combined model can capture spatio-temporal patterns. This not only speeds up evaluation in rice but also improves its accuracy, at least for rice. With various model architectures in deep learning, this study uses different temporal and spatial representations to produce better results. Unlike previous studies based on a single deep learning architecture, such as CNN [25], the new hybrid framework in this study has several advantages in plant phenotyping. Its performance exceeds those types but combines strengths more optimally for spatial and temporal modeling. Further research focuses on the best interaction between two data modes: RGB images and environmental parameters. This results in more accurate classification of rice growth stages. In the field of vertical farming, this study also combines pre-trained CNNs (VGG19, Xception, MobileNet, and ResNet) with trainable initial layers, global average pooling, and dense layers specific to growth stages. These changes enable the model to better identify useful spatial features and improve the classification precision of rice growth stages under controlled vertical farming conditions.

2.2 Data Acquisition and Construction

From the vertical farming system equipped with IoT sensors installed in strategic positions and high-resolution RGB cameras, two sets of data (multimodal data) were collected. On the one hand, daily image data were obtained from canopy-level RGB cameras. This data is recorded from day 0 to approximately 120 days after planting and displays the visual development of rice growth in images taken from a top or side view of the plants growing from the bottom up. On the other hand, sensor data consists of sequential, time-stamped measurements from fixed locations around the canopy. The six parameters measured were air

temperature T(t), relative humidity H(t), photosynthetically active radiation PAR(t), soil moisture SM(t), and water pH pH(t), all measured every ten minutes by a microcontroller-based data logger (NodeMCU ESP32). A total of 13,583 labeled images of rice plants collected from conventional rice fields were included in the dataset, each with a unique name for its growth stage. A parallel set containing 14,681 sensor data points was obtained from an IoT vertical farming model. Over a period of 122 days from initial planting, this device recorded pH, water levels, light intensity, ambient temperature, air humidity, soil moisture percentage, and soil pH. Data was collected from a vertical farming system equipped with strategically installed IoT sensors and high-resolution RGB cameras. The multimodal dataset consists of two primary sources: (1) image data, acquired daily from canopy-level RGB cameras from day 0 to approximately day 120 after planting, capturing the visual progression of rice growth, and (2) sensor data, consisting of continuous time-series measurements of environmental parameters, including air temperature T (t), relative humidity H (t), photosynthetically active radiation PAR (t), soil moisture SM (t), and water pH pH (t), recorded at fixed 10-minute intervals using microcontroller-based data loggers (NodeMCU ESP32). The available data set includes 13,583 rice plants marked as normal growing plants during the normal growth period, and the imaging blocks were taken in 8 rows. In addition, 14,681 sensor data points were collected from Internet of Things (IoT) devices embedded in the vertical farming model. These devices continuously monitored environmental factors affecting plant growth for up to 122 days after initial planting, when they stopped operating to report their findings: time, water pH level, typology, softness, magnitude of force, light intensity, spectrum range, and temperature at the boundary between air and soil. However, their reports this time were misleading. In foreign words without English equivalents, this was not very helpful before computers became commonplace about two decades ago — the original author mentions that it has been left as an unsolvable technical problem. The company also does not allow valid sources/validations for its reports. To prevent this problem, researchers must develop appropriate safeguards although if an unavoidable malfunction occurs, then everything will be lost, as it has already caused chaos in the research. The collected data needs to be checked for potential problems such as missing values, anomalies, and inconsistencies. This is an important pre-processing step before proceeding to the next stages of analysis and model formation. Each pair of synchronized image and sensor time-series data is labeled according to two criteria: Days After Planting (DAP), representing the temporal progression of plant development, and Growth Stage, categorizing each sample into one of eight standard rice growth phases Seedling, Early Vegetative, Maximum Tillering, Panicle Initiation, Booting, Grain Filling, Maturity, and Harvest with DAP serving as the primary reference [26], [27], [28].

Fig. 3. The growth stages of rice plants from seedling to harvest.

Table 1, the rice plants' growth stages, the physiological and environmental traits of each stage, suitable temperature/humidity regime, light condition and water/nutrient requirements (can be implemented in the proposed vertical rice cultivation system) for vertical rice farming.

Table 1. Environmental Requirements of Rice at Different Growth Stages [29], [30].

Growth Stage (Days After Planting)	Plant Characteristics	Optimal Conditions	Water & Nutrient Requirements
Nursery (0–14 DAP)	Height 5–15 cm, young green leaves	Temp: 25–30 °C; Humidity: 80–90%; Light not crucial	Saturated but not flooded, low nutrients
Early Vegetative (14–25 DAP)	1–3 tillers, 15–30 cm, bright green leaves	Temp: 25–32 °C (day), 20–25 °C (night); Humidity: 70–85%; Light: 10–12 h, medium intensity	Shallow water (2–5 cm), high nitrogen (N)
Maximum Tillering (25–45 DAP)	5–20 tillers, 30–55 cm, dark green leaves	Temp: 28–34 °C (day), 22–26 °C (night); Humidity: 65–80%; Light: 12–14 h, high intensity	±5 cm water depth, high nitrogen
Panicle Initiation (45–55 DAP)	Tillering stabilizes, 50–65 cm, dark green leaves	Temp: 25–30 °C; Humidity: 65–75%; Light: 12 h optimal	Stable water ±5–10 cm, need for P & K increases
Booting (55–70 DAP)	65–80 cm, grain initiation, green-yellow leaves	Temp: 25–32 °C; Humidity: 65–75%; Light: 12 h full	Stable water ±5–10 cm, high P & K demand
Grain Filling (70–95 DAP)	75–95 cm, yellowing leaves, grains filling	Temp: 24–28 °C; Humidity: 60–70%; Light: 12 h	Shallow water (2–5 cm), focus on potassium (K)
Maturation (95–115 DAP)	Yellow dominant leaves, grains hardening	Temp: 20–27 °C; Humidity: 50–65%; Light important for drying	Reduce water, initiate field drying
Pre-Harvest (>115 DAP)	Dry yellow leaves, hard grains	Temp: 20–27 °C; Humidity: 50–60%; Light not crucial	Stop irrigation 7–10 days before harvest

As shown in Table 1, each stage of rice plant growth requires a different environment for optimal growth. Leaf development and tillering can be observed from the seedling stage to the early vegetative stage, especially in shallow water conditions with high nitrogen availability. During the productive phase, especially in the early stages of panicle formation, phosphorus and potassium requirements are very high and depend on water availability above critical levels for optimal panicle development. During the grain-filling phase, potassium becomes the primary nutrient that determines grain size and weight as water availability decreases during the ripening-to-pre-harvest phase for drying and/or hardening. Understanding the unique nutrient requirements of plants at various stages of development is a key factor in achieving efficient resource use and increased crop production and quality, which are the foundation of a sustainable conventional agricultural system.

The database contains a collection of high-resolution images taken directly from multiple paddy fields at all eight normal growth stages (Seedling, Early Vegetative, Late Vegetative, Maximum Tillering, Panicle Initiation, Booting, Early Grain Filling, and Maturity/Harvest). Such labelled images lay a solid foundation for supervised deep learning, enabling models to grasp spatial features and the temporal progression of rice growth stages, which is instructive for classifying rice growth stages in real field practice.

Measurements were taken in an indoor vertical rice farm equipped with an IoT-connected environmental sensor network installed for continuous, autonomous operation. The sensor system collected information about the environment on a very detailed time scale over five months to create a data series with high temporal resolution (number of measurements per unit of time), which is ideal for temporal analysis and predictive modeling. Each sensor unit was designed to measure seven key

parameters related to microclimate regulation and plant physiological response, including irrigation water pH (ph_water), light intensity (light), air temperature (temperature), ambient relative humidity (humidity_ambient), soil moisture on an ADC scale (humidity_soil), soil pH (soil ph), and categorical variables [description] that indicate qualitative terms for measured variables, such as "moist" for soil.

Table 2. Descriptive statistics of environmental sensor variables

Variable	Mean	Standard Deviation (SD)	Min	Max
Water pH (ph_water)	7.47	1.11	4.50	18.50
Light (light)	554.91	1325.50	0.00	24385.83
Air Temp (°C) (temperature)	28.98	2.03	23.60	38.30
Ambient RH (%) (humidity_ambient)	83.74	9.63	51.00	100.00
Soil Moisture (ADC) (humidity_soil	1887.52	286.13	510.00	4095.00
Soil pH (ph_soil)	5.16	0.28	3.00	6.50

Table 2: Descriptive statistics of the environmental parameters in an indoor rice vertical farming system throughout a five-month monitoring period. The Dataset is a time series in minute level and includes the Real-time measurements from IoT-Based sensors such as PH of irrigation water (ph_water), soil pH(ph_soil), air temperature (temperature) (in Celsius), humidity_ambient (in Percentage), humidity of soil/pathway (soil_humidity) in ADC units and light intensity (light). The minimum, maximum, mean, and standard deviation values are provided as reported metrics to indicate the range of variability in microclimatic conditions within the controlled cultivation house. In Table 2 presents environmental parameters recorded in real-time, including: time, water pH, light intensity, ambient temperature, air humidity, soil moisture, soil pH, and soil condition description. Formally, let $X_i = \{X_i, S_i\ (t)\}$ represent the $i-th$ data sample, where I_i is the image tensor and $S_i\ (t)$ is the corresponding time-series sensor vector over time t. The dataset is thus represented as $\mathfrak{D} = \{(X_i y_i)\}\frac{N}{i=1}$ where $y_i \in = \{1, 2, 3 \ldots .8\}$ denotes the annotated growth stage. This labeled multimodal dataset provides a robust foundation for supervised learning tasks, including the classification of rice growth stages and modeling of plant development dynamics.

2.3 Data Processing

Data pre-processing was done independently in each data modality to maintain input fidelity and model preparedness. In the case of image data, all pixel values were divided by 255 to be normalized between 0 and 1, and all images were resized to a resolution of 224 × 224 for uniformity with pre-trained CNN architectures' input dimensions. During training, we applied data augmentation to the training set in order to improve generalization and reduce overfitting; we used the Augmentations library to implement random rotations (±450), horizontal/vertical flipping, zoom, and brightness/contrast transformations. For time-series sensor-level data, preprocessing included data cleaning (removal of missing values and outliers) using an appropriate interpolation scheme and feature normalization (μ is the mean and σ is the standard deviation) to stabilize training and speed up convergence.

The sensor sequences were then segmented and temporally aligned with the timestamps of the corresponding images, forming coherent multimodal input samples. $X_i = \{X_i, S_i\ (t)\}$ for the model, where I_i is the image tensor and $S_i\ (t)$ is the sensor vector over time t. Finally, the preprocessed dataset $\mathfrak{D} = \{(X_i y_i)\}\frac{N}{i=1}$ was stratified into three subsets to preserve proportional class distributions: Training Set (70%) to fit models, Validation Set (15%) for hyperparameter tuning and monitoring training process as well as testing, and Testing Set (15%), which we kept completely unseen until final evaluation. These extensive preprocessing steps result in spatial and temporal features being carefully normalized, augment, and synchronize spatial and temporal features, providing high-quality inputs for the hybrid CNN-LSTM architecture.

2.4 Hybrid Multi-Input Model Architecture

The core of the proposed framework is a Multi-Input Hybrid Deep Learning model that combines two different processing branches. The image data path takes canopy images as input and employs a pre-trained CNN as a spatial feature extractor. Four backbone architectures, VGG19, Xception, MobileNet, and ResNet50, are explored and compared. In all cases, the original classifier (top layers) is removed (include_top=False), yielding a high-level feature map. $\boldsymbol{F}_{Img}$ That encodes the visual characteristics of the rice canopy. A 2D Global Average Pooling layer follows the base model to convert the feature map into a compact 1D feature vector. $V_{Img} \in \mathbb{R}^d$ The sensor data path receives environmental time-series data as input and utilizes an LSTM layer to capture temporal dependencies. LSTM is specifically chosen for its capability to learn and retain long-term sequential patterns, modeling the temporal dynamics of environmental factors that influence plant growth. The resulting temporal feature vector is denoted as $V_{sensor} \in \mathbb{R}^k$.

The fusion and classification stage concatenates the spatial and temporal vectors into a multimodal representation. $V_{sensor} = [V_{Img};\ V_{sensor}]$, which is then passed through a series of fully connected (Dense) layers to learn complex non-linear correlations between visual and temporal features. The final output layer is a Dense layer with softmax activation, producing a probability distribution over the eight rice growth stages:

$$\hat{y} = softmax\left(W.V_{fusion} + b\right),\ \hat{y} \in \mathbb{R}^8 \qquad (1)$$

The final classification layer, defined by Equation 1, maps the combined multimodal feature representation to a probability distribution in the rice-plant growth class phase. The fusion feature matrix is obtained by combining spatial features extracted by the CNN backbone with temporal features modeled by the LSTM branch. This design allows simultaneous spatial-temporal learning from image data and environmental sensor data. The task is multi-class classification with a single label where each sample belongs to one of eight predefined rice-plant growth stages. Hence, the SoftMax activation function is used. Consequently, the output space can be succinctly described as $\hat{y} \in \mathbb{R}^8$, referring to the eight agronomic phenological categories used in this study. The bias vector b and weight matrix W in Equation (1) are not defined by hand. Rather, they are learned automatically during training using backpropagation. Cross-entropy loss is utilized to probabilistically optimize the parameters. This formulation is mathematically equivalent to a SoftMax-based probabilistic output. All networks can adaptively optimize decision boundaries by simply employing gradients computed from the propagated loss through the CNN backbone, LSTM branch, and classification head.

Where W and b are weights and biases that can also be learned from classification. This architecture facilitates the joint learning of spatial and temporal information, resulting in good classification accuracy for rice growth stages in IoT-connected vertical farming. The full process of this research flow is presented in Figure 4.

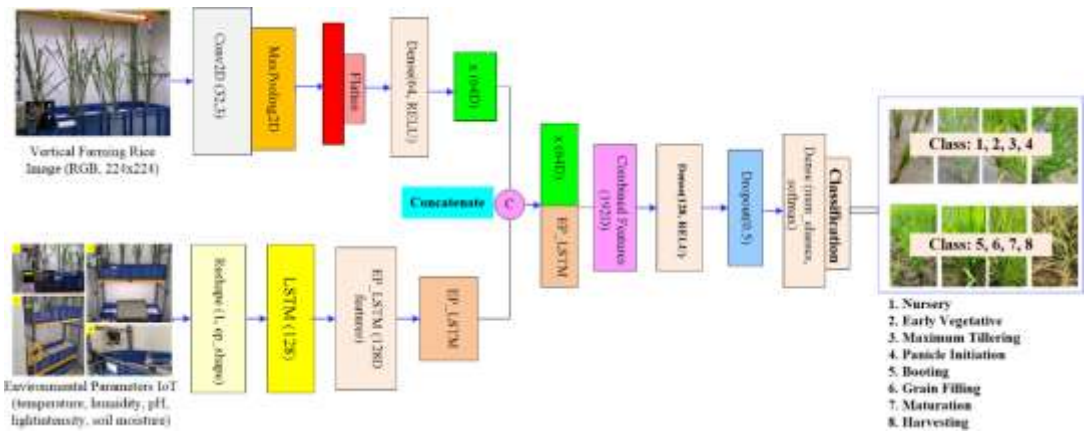


Fig. 4. Hybrid deep learning architecture for classifying eight phases of rice growth in an IoT-based vertical farming system

The Hybrid Deep Learning model (a cascading CNN with LSTMs) is depicted in Fig. 4, where both spatial and temporal processing are performed concurrently. The purpose of this design is to improve the labeling accuracy, or prediction tasks of multimodal data while aggregating different visual features in images and temporal patterns extracted through sensor measurements [31], [32]. As illustrated in the proposed architecture, Branch 1 (CNN branch) is designed to process RGB images with a spatial resolution of 224×224×3. The data is then convolved by multiple convolutional layers (Conv2D) and pooled by pooling layers (MaxPooling2D), which are flattened to be entered as input into a last ReLU-activated Dense layer, where stands for the activation function that extracts spatial characteristics. The output of branch 1 is output as the feature vector:

$$F(x) = \sigma(W * x + b) \quad (2)$$

Here is a function that models the convolutional layers of a CNN to produce the feature map as output. The convolution is the result of applying a kernel (W) to the input (x), adding a bias term (b), and then applying a nonlinearity (σ), which yields a nonlinear feature map. In this work, the activation function (σ) is ReLU [33].

Branch 2 (LSTM branch) receives environmental data from Internet of Things (IoT) sensors, including light intensity, temperature, humidity, and pH. They are arranged as fixed-length temporal sequences for LSTM processing. Then the data are passed through a layer of 128 LSTM units for encoding temporal information. Let be the input sensor matrix with sequence length T and where d=6 is the number of sensor features (water pH, light intensity, air temperature, air humidity, soil moisture and soil pH). The LSTM branch employs 128 hidden units to provide sufficient capacity to model multi-channel environmental time-series data, while maintaining computational efficiency and stability. Meanwhile, the CNN side generates a compact 64D feature vector to preserve useful spatial information and alleviate overfitting. The multi-ordinal nature of this dimensionality enables the balanced blending of temporal and spatial representations in a balanced manner in the multimodal framework. The 128-unit LSTM produces a 128-dimensional hidden state vector corresponding to the final time step, which serves as the temporal feature vector:

$$F_{LSTM} = LSTM(E) \in R^{128} \quad (3)$$

The outputs from both branches are subsequently fused using a concatenation operation to form a combined feature vector:

$$F_{combined} = Concat(F_{CNN}, F_{LSTM}) \in R^{192} \quad (4)$$

Here, the combined feature vector ($F_{combined}$) (F_{CNN}, F_{LSTM}) $\in R^{192}$ is obtained by concatenating the CNN feature vector $F_{CNN} \in R^{64}$ and the LSTM feature vector $F_{LSTM} = LSTM(E) \in R^{128}$. This concatenated vector is fed to the classification module, which includes one or more Dense layers, and is finally connected to a SoftMax layer to output the predicted probability distribution over rice growth stage classes. All parameters are end-to-end optimized using cross-entropy loss and the Adam optimizer, so backpropagation updates both the CNN and LSTM weights. This joint optimization enables efficient learning of spatial–temporal correlations and is one of the key novelties in this work. This structure enables the model to fully exploit spatial information in images and temporal sequences from environmental sensors, resulting in higher accuracy in classifying rice growth stages for IoT-enabled vertical farming systems. This hybrid CNN–LSTM model can be formulated into an algorithm as follows:

Hybrid Algorithm CNN (Multi-Backbone) + LSTM

Input: Image data **(*X*)**, Environmental Factor (EF) sequence data **(*Y*)**.
Output: Trained model weights for the optimal backbone, predicted sample visualizations, confusion matrix, and F1 score **(Θ, *b*)**.

1: **Initialize CNN backbone** CNN_b with pre-trained weights W^b_{CNN} (ImageNet) for each backbone (MobileNet, ResNet50, VGG19, Xception).
2: **Initialize the LSTM model** with weights W_{LSTM}
3: **Training Phase:**
Initialize model parameters Θ, $b = [W^b_{CNN} + W_{LSTM}]$.
Define the *CNN_b* architecture with convolutional and pooling layers.
Define the *LSTM* architecture with hidden states and gating mechanisms.

for epoch $\in \{1,\ldots,E\}$.
4: ***Forward pass:***
Extract spatial features from images: $F_{CNN} = CNN_b(X, W^b_{CNN})$.
Process features with LSTM: $F_{LSTM} = \text{LSTM}(F_{CNN}, W_{LSTM})$.
5: ***Loss calculation:***
Compute loss: $L = CrossEntropy(F_{LSTM}, Y)$.
6: ***Backpropagation:***
Compute gradients: $\nabla = Gradient(L, \Theta)$.
Update parameters with optimizer: $\Theta \leftarrow \Theta - \eta \nabla L$; → where η *is learning rate*
7: ***Testing Phase:***
Extract features with CNN: $F_{CNN} = \text{CNN}(X_{test}, W_{CNN})$.
Process features with LSTM: $F_{LSTM} = \text{LSTM}(F_{CNN}, W_{LSTM})$.
Compute evaluation metrics for each backbone $b \rightarrow M_b$ (e.g., accuracy, macro-F1).
8: **Model Selection:**
Determine the best-performing backbone $\mathbf{b^*} = argmax_b M_b$
9: **Prediction:** Use backbone ***b**** for final predictions.
10: **Return**: Optimal model parameters Θ_{b^*} and performance metrics M_b for all backbones.

Fig. 5. Hybrid multi-backbone CNN–LSTM algorithm for joint image and environmental data learning in rice growth stage prediction

The hybrid model processes spatial and temporal information simultaneously, improving the accuracy and classification of rice growth stages in IoT-based vertical farming systems.

2.1 Model Performance Evaluation

The performance of the proposed deep models was quantitatively evaluated using several commonly used metrics, including overall accuracy (OA), precision, recall, F-1 score, and Intersection over Union (IoU).

These metrics provide a comprehensive understanding of the model's classification capability by capturing different aspects of prediction quality as explained in the following equations [31]:

$$OA = \frac{\sum TP + \sum TN}{\sum TP + \sum TN + \sum FP + \sum FN} \quad (5)$$

$$Precision = \frac{\sum TP}{\sum TP + \sum FP} \quad (6)$$

$$Recall = \frac{\sum TP}{\sum TP + \sum FN} \quad (7)$$

$$F1 = 2\,x\,\frac{Precision\ x\ Recall}{Precision + Recall} \quad (8)$$

$$IoU = \frac{TP}{FP + TP + FN} \quad (9)$$

The classification performance of the proposed hybrid CNN–LSTM model was quantitatively evaluated based on a set of classic evaluation metrics: Overall Accuracy (OA), Precision, Recall, F1-score (F1), and

Intersection over Union (IoU), which can be calculated as in Equations (5) – (9). Accuracy (AC) is the ratio of correctly classified specimens, including true positives (TP) and true negatives (TN), to the total number of samples. Precision is the proportion of true positives among all predicted positives, and Recall is the proportion of true positives among all actual positives. As is shown in this figure, the F1-score, representing the harmonic mean of precision and recall, provides a well-balanced measure of the model's classification performance. Moreover, IoU serves as a tighter measure in terms of the amount of predicted class overlap with ground-truth classes and further complements other metrics by assessing the extent to which the two methods agree at both class-wise and whole-framework levels.

III. RESULTS

The results of the experiments indicate the efficiency of our hybrid CNN–LSTM architecture. Across all investigated backbone architectures, combining CNN with an LSTM outperformed the pure CNN baseline. Over the variants, VGG19–LSTM performed best overall, with a macro F1-score of ~ 0.96. In the early and late stages of growth, this approach achieved close-to-perfect (0.9) performance; for more complex intermediate stages, such as Panicle Initiation and Booting, it remained strong, achieving macro-average F1-scores above 0.89 across all growth stages. As shown in Table 3, with only 30 training epochs, the hybrid CNN–LSTM method is superior to the baseline model (CNN-only) for all backbone models (Mobile Net, ResNet50, VGG19, and Exceptions), which again demonstrates the effectiveness of combining temporal sequence learning with convolutional feature extraction. All precision, recall, F1-score, and IoU ratios reported in Table 3 are macro-averaged for all stages of rice plant growth. A row is a tuple (precision-macro avg., recall-macro avg., F1-score-macro avg., IoU) per class ratio. Performance measurements are computed separately for each growth stage and then averaged to obtain an overall estimate that gives equal weight to all classes. This enhancement is most significant in the context of rice growth-stage categorization, as subtle morphological differences across transition phases require models that capture both spatial and temporal dynamics. Evaluations show that multimodal fusion always brings improvements in accuracy, precision, recall, and F1-score overall growth stages. The VGG19–LSTM was the best choice of all proposed variants, since it had 0.96 as a macro F1-score. This implies that incorporating temporal sensor data provides complementary information for classification, especially during phenologically complex stages.

Table 3. Performance comparison of CNN-based and hybrid CNN–LSTM models for rice growth stage classification, reported as macro-averaged precision, recall, F1-score, and IoU across all growth stages.

No	Model Backbone	Overall Accuracy (OA)	Average			
			Precision	Recall	F1-Score	IoU
1	CNN	83.45%	0.85	0.77	0.73	0.58
2	MobileNet	91.57%	0.94	0.92	0.92	0.83
3	ResNet50	84.21%	0.82	0.84	0.82	0.69
4	VGG19	91.68%	0.92	0.92	0.92	0.83
5	Xception	90.55%	0.92	0.91	0.91	0.82
6	CNN+LSTM	89.47%	0.89	0.88	0.88	0.78
7	MobileNet+LSTM	93.02%	0.94	0.94	0.95	0.86
8	ResNet50+LSTM	92.48%	0.94	0.94	0.94	0.85
9	VGG19 + LSTM	96.64%	0.97	0.97	0.97	0.92
10	Xception+LSTM	95.20%	0.95	0.95	0.95	0.89

Table 3 presents the quantitative evaluation of each model architecture. The superior performance of the hybrid CNN–LSTM models compared to their CNN-only counterparts in all evaluation metrics is an indicator of the benefit gained when integrating temporal (environmental: via LSTM) and spatial (visual: through CNN) information. From the hybrid structures, VGG19–LSTM performed the best in all respects: 96% of accuracy, macro F1-score 0.96, and high IoUs (up to 0.92 on key classes). Mobile Net and LSTM with ResNet50 and LSTM also achieved very high-performance levels, with accuracies of up to 94%. The Xception model with LSTM is reported to have an accuracy of 93%. This confirms that the advantages of the hybrid framework generalize well across different CNN backbones. On the other hand, single CNNs achieved lower accuracy, ranging from 75% (plain CNN) to 92% (Mobile Net CNN). The difference in performance is particularly noticeable when comparing VGG19 (88%) to its hybridized variant, which achieves 96%, corresponding to an 8% performance increase. These findings show that spatial visual clues alone are insufficient to differentiate growth stages with similar shapes, and a hybrid approach may help in effectively integrating multimodal spatial and temporal information. The detailed classification results for all tested architectures across the eight growth stages in rice (Seedling, Early Vegetative, Maximum Tillering, Panicle Initiation, Booting, Early Grain Filling Maturity, and Harvest) are listed in Table 4. The reported stats include Overall Accuracy (OA), Precision, Recall, F1-score, and IoU, which provide an overall assessment of the model.

Table 4. Summary of evaluation metrics for all model architectures

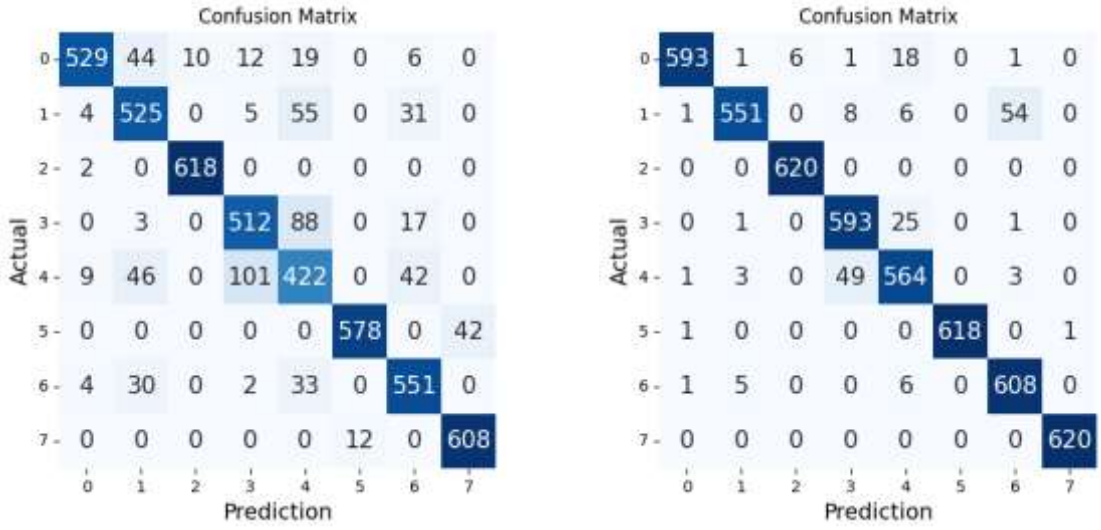
Class	No	Overall Accuracy	Precision	Recall	F1-Score	IoU
Seedling	1	0.82	0.99	0.92	0.96	0.92
	2	0.91	0.99	0.97	0.98	0.96
	3	0.79	0.90	0.86	0.88	0.79
	4	0.94	0.97	0.95	0.96	0.93
	5	0.9	0.98	0.96	0.97	0.94
	6	0.92	0.99	1	0.99	0.98
	7	0.95	1	1	1	1
	8	0.93	0.99	0.99	0.99	0.98
	9	0.97	1	0.99	0.99	0.99
	10	0.96	1	0.99	0.99	0.99
Early Vegetative	1	0.82	0.92	0.97	0.95	0.90
	2	0.91	0.96	0.99	0.98	0.96
	3	0.79	0.86	0.945	0.90	0.82
	4	0.94	0.95	0.98	0.96	0.93
	5	0.9	0.96	0.98	0.97	0.94
	6	0.92	0.98	0.98	0.98	0.96
	7	0.95	1	1	1	1
	8	0.93	0.99	1	0.99	0.99
	9	0.97	1	1	1	1
	10	0.96	1	0.99	0.99	0.99
Maximum Tillering	1	0.82	0.91	0.71	0.80	0.67
	2	0.91	0.96	0.83	0.89	0.81
	3	0.79	0.70	0.78	0.74	0.59
	4	0.94	0.94	0.82	0.87	0.78

Class	No	Overall Accuracy	Precision	Recall	F1-Score	IoU
	5	0.9	0.89	0.85	0.87	0.78
	6	0.92	0.96	0.80	0.87	0.77
	7	0.95	0.95	0.90	0.93	0.87
	8	0.93	0.95	0.90	0.93	0.86
	9	0.97	0.98	0.94	0.96	0.92
	10	0.96	0.94	0.88	0.91	0.84
Panicle Initiation	1	0.82	0.81	0.88	0.84	0.73
	2	0.91	0.93	0.92	0.93	0.87
	3	0.79	0.85	0.76	0.80	0.66
	4	0.94	0.88	0.89	0.88	0.79
	5	0.9	0.93	0.92	0.92	0.86
	6	0.92	0.94	0.92	0.93	0.87
	7	0.95	0.95	0.92	0.94	0.89
	8	0.93	0.94	0.94	0.94	0.89
	9	0.97	0.91	0.97	0.94	0.89
	10	0.96	0.95	0.92	0.94	0.88
Booting	1	0.82	0.78	0.40	0.53	0.36
	2	0.91	0.844	0.92	0.88	0.79
	3	0.79	0.76	0.72	0.74	0.59
	4	0.94	0.83	0.87	0.85	0.74
	5	0.9	0.84	0.84	0.84	0.73
	6	0.92	0.75	0.95	0.84	0.72
	7	0.95	0.88	0.90	0.89	0.80
	8	0.93	0.96	0.91	0.94	0.88
	9	0.97	0.99	0.89	0.94	0.89
	10	0.96	0.90	0.85	0.87	0.78
Early Grain Filling	1	0.82	0.85	0.15	0.26	0.15
	2	0.91	0.81	0.75	0.78	0.64
	3	0.79	0.58	0.71	0.64	0.47
	4	0.94	0.70	0.69	0.69	0.53
	5	0.9	0.80	0.75	0.77	0.63
	6	0.92	0.80	0.46	0.58	0.41
	7	0.95	0.82	0.84	0.83	0.71
	8	0.93	0.87	0.86	0.87	0.77
	9	0.97	0.88	0.89	0.89	0.80
	10	0.96	0.77	0.89	0.83	0.71
Maturity	1	0.82	0.40	0.98	0.57	0.40
	2	0.91	0.85	0.94	0.89	0.81
	3	0.79	0.89	0.69	0.78	0.63
	4	0.94	0.82	0.86	0.84	0.72

Class	No	Overall Accuracy	Precision	Recall	F1-Score	IoU
	5	0.9	0.86	0.94	0.90	0.82
	6	0.92	0.67	0.88	0.76	0.62
	7	0.95	0.91	0.94	0.93	0.87
	8	0.93	0.87	0.96	0.91	0.84
	9	0.97	0.90	0.94	0.92	0.85
	10	0.96	0.92	0.92	0.92	0.86
Harvest	1	0.82	0.98	1	0.99	0.98
	2	0.91	0.98	1	0.99	0.98
	3	0.79	0.95	0.96	0.96	0.92
	4	0.94	0.97	1	0.98	0.97
	5	0.9	0.98	1	0.99	0.98
	6	0.92	0.99	0.99	0.99	0.98
	7	0.95	0.99	1	0.99	0.99
	8	0.93	0.97	0.98	0.9	0.9
	9	0.97	0.98	1	0.99	0.98
	10	0.96	0.98	1	0.99	0.98

*1. CNN; 2. MobileNet; 3. ResNet50; 4. VGG19; 5. Xception; 6. CNN+LSTM; 7. MobileNet+LSTM; 8. ResNet50+LSTM; 9. VGG19+LSTM; 10. Xception+LSTM

The results obtained in the experiments on all model variations demonstrate that our tensor-based CNN–LSTM framework outperforms CNN-only baselines. CNN-only showed moderate performance, with the best classification accuracy ranging from 75% (CNN) to 91% (CNN-Xception). These models frequently failed at the intermediate growth stage (e.g., Booting, Grain Filling) to some extent, due solely to limitations in visual features. CNN–LSTM models performed better than CNN-only networks, achieving 88% and 96% accuracy (CNN-LSTM and VGG19-LSTM, respectively). This improvement was particularly evident in the difficult stages (e.g., at Booting and Grain Filling), where the temporal sensor's characteristics complemented those based on visual features. In the case of hybrids, VGG19-LSTM achieved the best results with 96% accuracy, a macro-F1 score of 0.96, and an IoU of 0.92, making it the best-performing backbone.MobileNet-LSTM and ResNet50-LSTM also delivered strong results (accuracy of 94%), confirming the strength of our hybrid approach across low-depth and deep architectures. We observe that the margin between 88% and 96% for VGG19-only vs. VGG19-LSTM is substantial. This 8% increase shows that it is very important to include environmental factor signals from temporal data in order to distinguish growth stages that are visually difficult to differentiate. To provide a more detailed description of the classification character, confusion matrix plots were generated for the best models (Figure 4). These visualizations allow a straightforward comparison between the CNN-only baseline (VGG19) and its hybrid extension with LSTM. Comparing both panels, we clearly see the potential to integrate temporal features in order to reduce misclassification, especially at similar early stages.



(a) (b)

Fig. 6. Confusion matrices with VGG19 backbone: CNN-only (a) vs. CNN–LSTM (b), where the hybrid model reduces misclassification and improves accuracy.

As presented in Table 3 and Table 4, the VGG19–LSTM hybrid performed well (the best overall accuracy as well as being more balanced across metrics). For example, Figure 4 shows the confusion matrices for the VGG19 backbone, contrasting the CNN-only and CNN–LSTM models. The reduced misclassification between morphologically close stages is visualized, thereby validating the quantitative improvements reported in Tables 3 and 4.

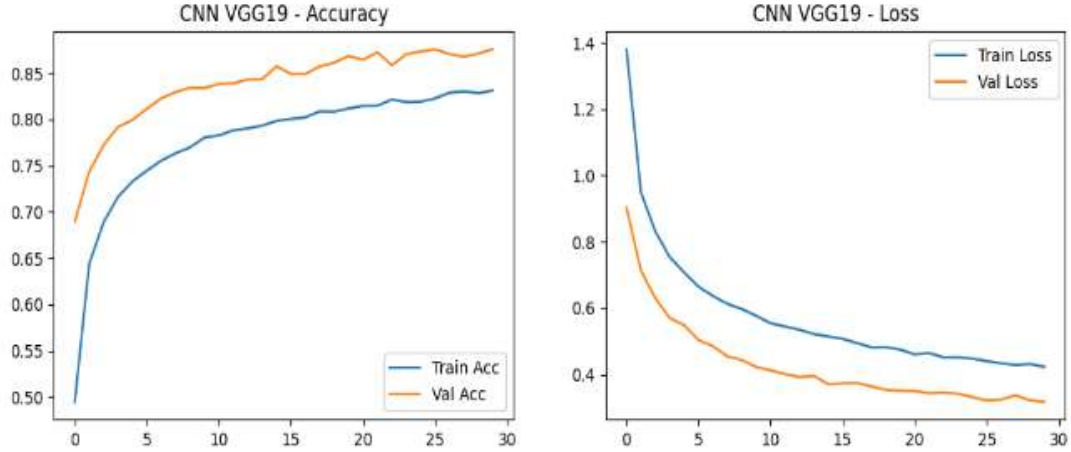


Fig. 7. Training and validation CNN VGG19 baseline model

Performance of baseline CNN VGG19 is shown in Fig. 6. There is a consistent linear increase of the training and validation accuracy towards higher epochs, while validation still converges at around 0.86. The trained loss and the validation losses are all down, and the two curves still maintain a significant gap. These results suggest that, despite the model effectively learning important characteristics, the moderate validation accuracy is due to slight underfitting, indicating only a limited ability to generalize.

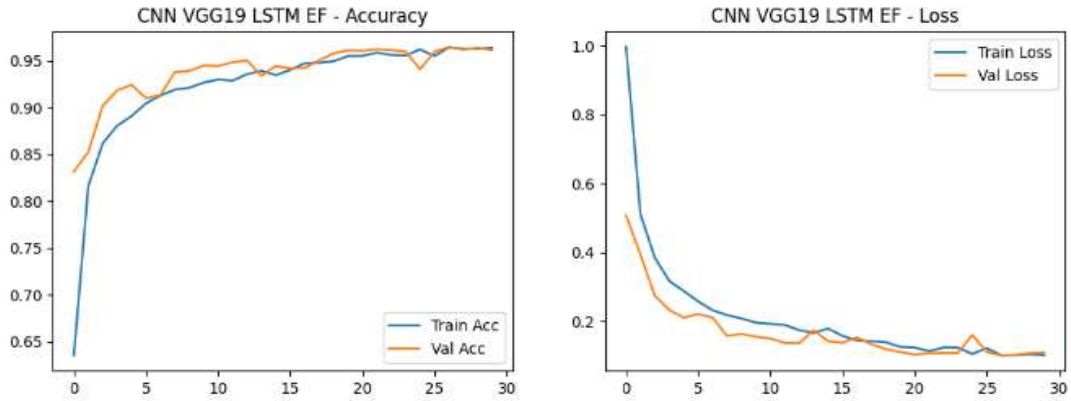


Fig. 8. Training and validation CNN VGG19–LSTM hybrid model

In contrast, Fig. 8 illustrates the behavior of the VGG19–LSTM hybrid CNN model used in this study, which yields better results. Training and validation accuracy exceed 0.95 with close-fitting curves, indicating better accuracy and generalization capabilities compared to the baseline model. Similarly, training and validation losses decreased to 0.1–0.2, with a very similar growth trend. These results suggest that integrating an LSTM helps the model capture information across time-steps, thereby lowering classification error for rice transition stages with small differences.

Qualitative prediction examples. Along with the tables and confusion matrices, we present qualitative prediction examples to give an idea of how well our model works in practice. Example input and ground truth/predicted class images are shown in Fig. 9 (which also shows the correct classification, including typical misclassified cases). The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

Fig. 9. Qualitative prediction results of the proposed VGG19+LSTM model

Fig. 9, qualitative prediction of the proposed VGG19+LSTM model (Overall Accuracy = 96.64%, F1-Score = 0.97, IoU = 0.92). The top row shows correctly classified samples across four rice growth stages (Seedling, Early Vegetative, Panicle Initiation, and Harvest), whereas the bottom row shows misclassifications such as Booting as Maturity, Maximum Tillering as Early Vegetative, and Maturity as Booting. This visual presentation demonstrated the model's capability to represent stages of rice plant growth and, at the same time, its limitation in discriminating morphologically similar stages.

IV. DISCUSSION

The experimental results demonstrate the effectiveness of fusing temporal sensor information with spatial visual features for rice growth-stage identification in IoT-oriented vertical farming. For all tested backbones, the hybrid CNN–LSTM architectures achieved superior performance compared to their CNN-only counterparts, indicating the significance of sequential environment information that cannot be fully represented by images alone. This provides supporting evidence that multimodal fusion is a fundamental technique for fine-grained phenological classification.

One of the major drawbacks of CNN-only models was frequent confusion between morphologically similar stages, for instance, Booting vs Maturation, and Early Grain Filling vs Panicle Initiation. This limitation is manifested quantitatively through lower recall and F1-score for these phases, as well as qualitatively in the confusion matrices. These data indicate that visual features are necessary but not sufficient for fine discrimination of phenological processes. These errors were reduced in the proposed hybrid CNN–LSTM models by leveraging temporal relations (e.g., cumulative thermal exposure and humidity cycles), which provide contextual information required to discriminate stages with similar visual pattern features.

In addition, the simpler models, such as VGG19 and LSTM, also performed well, with OA of 96.64%, F1 score of 0.97, and IoU of 0.92, among other methods implemented in this study. These models also showed that their sulcus segmentation performance was superior to that of classical deep learning architectures. Their high performance lies in the deep hierarchical feature extraction of VGG19, which synergizes with motion information to produce highly discriminative spatio-temporal representations. In addition, MobileNet and LSTM, as well as ResNet50 + LSTM, achieved competitive results with an accuracy (OA) of over 93%, indicating that the improvement in performance was not driven by a single backbone but by hybrid models. These results further support the use of CNN-LSTM fusion, a generally applicable approach for plant growth stage classification.

The VGG19 and LSTM models are robust because their training paths are stable: both the training and validation loss curves converge without signs of overfitting. Suffice it to say that for a smart farming system to function in the real world, it must be stable enough to support common operations such as automatic irrigation, nutrient supply, and harvest timing. Further qualitative evaluation confirms that the model successfully handles extreme cases, but residual errors are observed during transition periods with subtle phenotypic changes and strong temporal signals.

This study provides results showing that the hybrid CNN and LSTM model, tested on spatial and temporal data from rice plants, is a reliable and efficient technique for monitoring details of rice growth stages under controlled conditions. The next step for future research is to develop this model in data modes such as hyperspectral imagers, soil nutrient sensors, or real-time weather inputs. Moreover, it is possible that an attention mechanism can be added for dynamically weighting the contributions of spatial features and temporal features to better performance of the model in real-world applications or interpretation.

V. CONCLUSION

The results of this work indicate that a convolutional neural network combined with a long short-term memory-based deep learning method can be used to classify rice growth stages in an IoT-based vertical farm, and that combining RGB canopy image features with temporal environmental sensor information reliably capture the spatio-temporal dynamics of plant growth. Experimental results show that the VGG19 and LSTM variants exhibit peak performance (96.64% accuracy, 0.97 macro-F1 score, and 0.92 IoU),

significantly better than the CNN-only baseline and other backbones (MobileNet, ResNet-50, VGG-19, Xception), while the multimodal hybrid approach systematically reduces classification errors, especially in visually similar stages such as Booting and Grain Filling. These results demonstrate that combining camera-based sequences and sensors can perform detailed phenological classification and provide automated, customized monitoring for irrigation scheduling, nutrient management, or harvest planning in vertical farms, ultimately contributing to sustainable vertical farming. In addition to high accuracy performance, the proposed approach has the potential to develop solutions that can be applied in real-time smart agriculture, in order to promote precision agriculture in scenarios with limited land and resources, while opening up opportunities for future research in integrating other modes such as hyperspectral images or soil nutrient sensors, as well as attention mechanisms to achieve improved performance and insights. Overall, this contribution provides new and practical insights into the development of smart farming systems and global food security.

Declarations

Author contribution. The contribution or credit of the author must be stated in this section.
Funding statement. The funding agency should be written in full, followed by the grant number in square brackets and year.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

REFERENCES

[1] R. Sikka, D. Pal Singh, M. K. Sharma, and A. Ojha, "Advancing Agriculture in Smart Cities: Renewable Energy and Artificial Intelligence-Powered IoT," E3S Web Conf., vol. 540, p. 13010, 2024, doi: 10.1051/e3sconf/202454013010.

[2] M. Abdulla and A. Marhoon, "Agriculture based on Internet of Things and Deep Learning," Iraqi J. Electr. Electron. Eng., vol. 18, no. 2, pp. 1–8, 2022, doi: 10.37917/ijeee.18.2.1.

[3] M. Woźniak and M. F. Ijaz, "Editorial: Recent advances in big data, machine, and deep learning for precision agriculture," Front. Plant Sci., vol. 15, 2024, doi: 10.3389/fpls.2024.1367538.

[4] A. Kempelis, I. Polaka, A. Romanovs, and A. Patlins, "Computer Vision and Machine Learning-Based Predictive Analysis for Urban Agricultural Systems," Futur. Internet, vol. 16, no. 2, p. 44, 2024, doi: 10.3390/fi16020044.

[5] N. E. Korres, J. K. Norsworthy, N. R. Burgos, and D. M. Oosterhuis, "Temperature and drought impacts on rice production: An agronomic perspective regarding short- and long-term adaptation measures," Water Resources and Rural Development. Accessed: Jul. 27, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S2212608216300389?utm_source=chatgpt.com

[6] S. F. ul Islam, A. de Neergaard, B. O. Sander, L. S. Jensen, R. Wassmann, and J. W. van Groenigen, "Reducing greenhouse gas emissions and grain arsenic and lead levels without compromising yield in organically produced rice," Agriculture, Ecosystems and Environment.

[7] X. Yao et al., "Optimizing water and nitrogen management to balance greenhouse gas emissions and yield in Chinese rice paddies," F. Crop. Res., vol. 319, p. 109621, Dec. 2024, doi: 10.1016/j.fcr.2024.109621.

[8] A. Faraji, A. Hosseini, M. Z. Kermani, N. Mashatan, and S. Ardestani, "Vertical Farming: Vertical Farming; an Innovative Agricultural Method to the Urban and Environmentally Sustainable Development," J. Innov. Sustain. RISUS, vol. 14, no. 3, pp. 166–181, 2023, doi: 10.23925/2179-3565.2023v14i3p166-181.

[9] I. Umarie, Oktarina, W. Widiarti, B. Suroso, M. Hazmi, and F. Podesta, "Potential and Challenges of Developing Hydroponic Rice Cultivation as Vertical Farming in Major Indonesian Cities," Int. J. Agribus. Sustain. Dev. Res., vol. 1, no. 4, p. 15, 2025, [Online]. Available: https://gscjournal.com/IJASDR/article/view/50

[10] R. R. A. Siregar, K. B. Seminar, S. Wahjuni, and E. Santosa, "Vertical Farming Perspectives in Support

of Precision Agriculture Using Artificial Intelligence: A Review," Computers, vol. 11, no. 9, 2022, doi: 10.3390/computers11090135.
[11] A. K. Podder et al., "IoT based smart agrotech system for verification of Urban farming parameters," Microprocess. Microsyst., vol. 82, p. 104025, 2021, doi: 10.1016/j.micpro.2021.104025.
[12] R. Al-Qudah, M. Almuhajri, and C. Y. Suen, "Unveiling the potential of sustainable agriculture: A comprehensive survey on the advancement of AI and sensory data for smart greenhouses," Comput. Electron. Agric., vol. 229, p. 109721, 2025, doi: 10.1016/j.compag.2024.109721.
[13] S. B. Dhal and D. Kar, "Transforming Agricultural Productivity with AI-Driven Forecasting: Innovations in Food Security and Supply Chain Optimization," Forecasting, vol. 6, no. 4, pp. 925–951, 2024, doi: 10.3390/forecast6040046.
[14] S. Atalla et al., "IoT-Enabled Precision Agriculture: Developing an Ecosystem for Optimized Crop Management," Inf., vol. 14, no. 4, pp. 1–23, 2023, doi: 10.3390/info14040205.
[15] M. Abdalla, O. A. Mohamed, and E. M. Azmi, "Adaptive Learning Model for Detecting Wheat Diseases," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 5, 2024, doi: 10.14569/ijacsa.2024.01505130.
[16] Y. M. Qin, Y. H. Tu, T. Li, Y. Ni, R. F. Wang, and H. Wang, "Deep Learning for Sustainable Agriculture: A Systematic Review on Applications in Lettuce Cultivation," 2025, Multidisciplinary Digital Publishing Institute. doi: 10.3390/su17073190.
[17] P. K. Sethy, N. K. Barpanda, A. K. Rath, and S. C. Rajpoot, "Rice (Oryza Sativa) panicle blast grading using support vector machine based on deep features of small CNN," Arch. Phytopathol. Plant Prot., vol. 54, no. 15–16, pp. 1001–1013, 2021, doi: 10.1080/03235408.2020.1869386.
[18] L. Yuchen, Z. Xusen, L. Jiali, and Z. Xueting, "Rice Leaf Nitrogen Deficiency Image Classification Model Based on CNN," Acad. J. Comput. Inf. Sci., vol. 6, no. 2, pp. 16–22, 2023, doi: 10.25236/ajcis.2023.060203.
[19] M. El Sakka, M. Ivanovici, L. Chaari, and J. Mothe, "A Review of CNN Applications in Smart Agriculture Using Multimodal Data," 2025, Multidisciplinary Digital Publishing Institute. doi: 10.3390/s25020472.
[20] V. Chaowalittawin, W. Krungseanmuang, P. Sathaporn, F. Morita, T. Archevapanich, and B. Purahong, "Banana quality classification using lightweight CNN model with microservice integration system," Eng. Appl. Sci. Res., vol. 52, no. 4, pp. 430–438, Jul. 2025, doi: 10.14456/easr.2025.38.
[21] K. Bansal et al., "Evolving CNN with Paddy Field Algorithm for Geographical Landmark Recognition," Electron., vol. 11, no. 7, 2022, doi: 10.3390/electronics11071075.
[22] A. Kamilaris and F. X. Prenafeta-Boldú, "A review of the use of convolutional neural networks in agriculture," 2018. doi: 10.1017/S0021859618000436.
[23] M. Nautiyal et al., "Revolutionizing agriculture: A comprehensive review on artificial intelligence applications in enhancing properties of agricultural produce," 2025, Elsevier BV. doi: 10.1016/j.fochx.2025.102748.
[24] D. Tamayo-Vera, X. Wang, and M. Mesbah, "A Review of Machine Learning Techniques in Agroclimatic Studies," 2024, Multidisciplinary Digital Publishing Institute. doi: 10.3390/agriculture14030481.
[25] J. Philip, "International Journal For Innovative Research In Multidisciplinary Field CNN-LSTM Hybrid Deep Learning Model for Remaining Useful Life Estimation," Int. J. Innov. Res. Multidiscip. F., vol. 10, no. Special Issue-54, pp. 38–50, 2024, [Online]. Available: https://www.ijirmf.com
[26] R. T. C. Sheng, Y. H. Huang, P. C. Chan, S. A. Bhat, Y. C. Wu, and N. F. Huang, "Rice Growth Stage Classification via RF-Based Machine Learning and Image Processing," Agric., vol. 12, no. 12, pp. 1–23, 2022, doi: 10.3390/agriculture12122137.
[27] S. K. De Datta, Principles and practices of rice production, vol. 17. A WILEY-Interscience Publication, 1385.
[28] Y. Shouichi, "Fundamentals of rice crop science," pp. 167–186, 2021.
[29] Z. Liu et al., "Effects of Temperature Fluctuations on the Growth Cycle of Rice," Agric., vol. 15, no. 1, pp. 1–13, 2025, doi: 10.3390/agriculture15010099.
[30] S. Echeverría-Progulakis et al., "Climate change mitigation through irrigation strategies during rice growing season is off-set in fallow season," J. Environ. Manage., vol. 380, no. April, 2025, doi:

10.1016/j.jenvman.2025.125060.
[31] O. Elsherbiny, L. Zhou, Y. He, and Z. Qiu, "A novel hybrid deep network for diagnosing water status in wheat crop using IoT-based multimodal data," Comput. Electron. Agric., vol. 203, no. September, 2022, doi: 10.1016/j.compag.2022.107453.
[32] A. Nazir et al., "A deep learning-based novel hybrid CNN-LSTM architecture for efficient detection of threats in the IoT ecosystem," Ain Shams Eng. J., vol. 15, no. 7, 2024, doi: 10.1016/j.asej.2024.102777.
[33] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," 2015. doi: 10.1038/nature14539.