Journal of Agricultural Sciences (JAS) 32(4):1030-1050
DOI: 10.15832/ankutbd.1898342

J Agr Sci-Tarim Bili
e-ISSN: 2148-9297
jas.ankara.edu.tr

Early Detection of Apple Scab in Orchards and Mildew in Vineyards Using A Novel Deep Learning Model and Multispectral Imagery

Batuhan Asiroglu[a], Aydin Kagan Turkoglu[b], Fatih Sinan Bir[c], Sibel Senan[a*]

[a]*Department of Computer Engineering, Faculty of Engineering, Istanbul University-Cerrahpasa, Istanbul, TÜRKIYE*

[b]*Department of Computer Engineering, Faculty of Electrical and Electronics, Yildiz Technical University, Istanbul, TÜRKIYE*

[c]*Department of R&D, B Tech Analytics Technology Inc., Istanbul, TÜRKIYE*

ARTICLE INFO
Research Article
Corresponding Author: Sibel Senan, E-mail: ssenan@iuc.edu.tr
Received: 26 February 2026 / Revised: 24 April 2026 / Accepted: 13 May 2026 / Online: 12 October 2026

Cite this article
Asiroglu B, Turkoglu A K, Bir F S, Senan S (2026). Early Detection of Apple Scab in Orchards and Mildew in Vineyards Using A Novel Deep Learning Model and Multispectral Imagery. *Journal of Agricultural Sciences (Tarim Bilimleri Dergisi),* 32(4):1030-1050. DOI: 10.15832/ankutbd.1898342

ABSTRACT
Agricultural diseases pose a significant threat to crop productivity by causing substantial damage to plant tissues and reducing yield. Therefore, accurate and early detection of plant diseases is critical for sustainable agricultural practices. In this study, a novel hybrid deep learning model is proposed for the detection and segmentation of plant diseases, specifically apple scab in orchards and mildew in vineyards. The proposed architecture integrates a U-Net framework with ConvLSTM to jointly model spatial, spectral, and temporal information from multispectral drone imagery. Unlike conventional U-Net + LSTM approaches, the model is adapted to process five-channel multispectral data and explicitly capture temporal dependencies across sequential observations, enabling improved detection of early-stage disease patterns. Multispectral time-series data were collected using a DJI Phantom 4 Multispectral drone over agricultural fields in Tekirdag, Turkey. The proposed model achieved an average Intersection over Union (IoU) of 82.05% for apple orchards and 79.67% for vineyards, demonstrating robust segmentation performance under real-world conditions. These results highlight the effectiveness of integrating multispectral and temporal information within a unified deep learning framework for accurate and timely plant disease detection in precision agriculture.

Keywords: Precision Agriculture, Plant Disease Detection, Convolutional LSTM, U-Net Architecture, Multispectral Time-Series, Spatiotemporal Analysis

1. Introduction

The productivity and quality of agricultural products are significantly influenced by environmental factors. Biotic and abiotic stressors, such as climatic anomalies, pests, and plant pathogens, can lead to substantial losses in agricultural production (Savary et al. 2019; Shahi et al. 2023). To mitigate these risks, pesticides are widely used; however, their excessive and continuous application increases production costs and poses serious risks to environmental and human health. In particular, pesticide residues in food products represent a major concern for consumer safety (Hua & Liu 2024). In addition, climate change manifested through rising temperatures and irregular rainfall patterns further accelerates the spread of plant diseases and pests, increasing the vulnerability of agricultural systems (Bebber et al. 2020). Therefore, early detection of plant diseases is critical for reducing economic losses, minimizing pesticide use, and ensuring sustainable agricultural practices (Mahlein 2016).

Traditionally, plant disease monitoring has relied on visual inspection. Although manual assessments can provide valuable insights, they are labor intensive, time consuming, and require expert knowledge, limiting their scalability and effectiveness for early-stage detection. Consequently, automated disease detection systems have become increasingly important in modern agriculture. However, many conventional approaches, including Normalized Difference Vegetation Index (NDVI) based methods (Avotins et al. 2020; Stamford et al. 2023), and Convolutional Neural Network (CNN)-based models (Shahi et al. 2023; De Silva & Brown 2022), primarily detect diseases after visible symptoms appear. This limitation reduces their effectiveness in identifying early-stage physiological changes and delays timely intervention.

To address these challenges, recent studies have explored hybrid deep learning architectures that combine CNNs with Long Short-Term Memory (LSTM) networks to analyze sequential image data (De Silva & Brown 2022; Cao et al. 2025). While CNNs are effective at extracting spatial features, LSTM networks capture temporal dependencies, making such hybrid models suitable for time-series analysis. In parallel, U-Net has emerged as a powerful architecture for image segmentation tasks, particularly in biomedical and agricultural applications. With its encoder–decoder structure and skip connections, U-Net enables precise pixel-level classification even with limited labeled data (Ronneberger et al. 2015; Chen et al. 2020). Building upon these

strengths, integrating temporal modeling capabilities into U-Net based frameworks offers a promising direction for improved disease detection. However, existing studies often fail to jointly exploit multispectral information and temporal dynamics within a unified segmentation framework, particularly for early stage plant disease detection. This limitation motivates the need for models that can effectively integrate spatial, spectral, and temporal features in a single architecture.

Multispectral imaging further enhances early disease detection by capturing information beyond the visible spectrum. In particular, Near-Infrared (NIR) and Red Edge bands can reveal disease symptoms earlier than RGB imagery alone, highlighting subtle physiological changes in plants (Bleasdale & Whyatt 2025). In this study, multispectral data were collected using a DJI Phantom 4 Multispectral Drone equipped with five spectral bands (Blue, Green, Red, Red Edge, and NIR) over apple orchards and vineyards in Tekirdag, Turkey. Due to operational constraints, data were acquired at irregular intervals, covering 29 weeks for apple orchards and 25 weeks for vineyards. The dataset consists of 3625 TIF samples for apple orchards (29 weeks × 5 bands × 25 grids) and 5625 TIF samples for vineyards (25 weeks × 5 bands × 45 grids). These time-series data enable the analysis of plant health dynamics across different growth stages.

In this study, we propose a novel hybrid deep learning framework for early plant disease detection, specifically targeting apple scab and vineyard mildew. The proposed architecture integrates LSTM layers into a customized U-Net framework to jointly model spatial, spectral, and temporal information from multispectral time-series data. Unlike conventional approaches, the model is specifically designed to leverage multispectral inputs together with temporal dynamics within a unified segmentation framework, enabling more effective detection of early-stage disease patterns. Experimental results demonstrate that the proposed model achieves an average Intersection over Union (IoU) of 82.05% for the apple dataset and 79.67% for the vineyard dataset. In addition, the model attains high classification performance, with categorical accuracy values of 96.68% and 92.16% for apple and vineyard datasets, respectively. These results demonstrate the effectiveness of integrating multispectral time-series data with hybrid deep learning architectures. The proposed approach provides a scalable and robust solution for precision agriculture by enabling accurate and timely disease detection through temporally informed segmentation.

2. Related Works

Remote sensing technologies combined with artificial intelligence have become essential tools for early detection and monitoring of plant diseases in precision agriculture. Recent advances in Drone imaging systems, particularly multispectral and hyperspectral sensors, provide improved spatial and spectral resolution compared to conventional RGB imagery. Previous studies have employed various CNN, LSTM, and U-Net architectures for disease detection, but most approaches focus on image-level classification and do not leverage temporal sequences or pixel-level segmentation for precise mapping.

Unlike conventional approaches that mainly use RGB images and classify entire images as healthy or diseased, our method performs pixel-level semantic segmentation using five-band multispectral imagery captured by drones (Blue, Green, Red, Red Edge, NIR). By integrating LSTM layers into a U-Net backbone, the proposed model exploits temporal sequences over repeated Drone flights, enabling detection of symptom development and progression across time, which is rarely addressed in prior studies. Table 1 presents a combined comparison of related works, summarizing their input data, targeted diseases, model architectures, data types, tasks, performance values and setting details with limitations.

Representative studies in the literature demonstrate common methodological limitations:

- Verma & Dubey (2021), Gnanasaravanan et al. (2021) and Farooqui et al. (2023) applied LSTM-based models to RGB data, achieving high classification accuracy while not providing pixel-level spatial localization.
- Madeira et al. (2024) and Sundravadivelu et al. (2023) combined CNNs with satellite or drone imagery for coarse disease detection but did not implement pixel-level segmentation.
- Ashwini & Sellam (2024), Bleasdale (2024), Rous et al. (2023), Zhang et al. (2025), and Zhang et al. (2025)used RGB or hyperspectral data in controlled settings, limiting applicability for early-stage detection under field conditions.
- Mydhili et al. (2025) utilized RGB data to achieve pixel-level spatial localization using close-range fruit images, but the reliance on controlled, close-range acquisition limits its applicability for early-stage disease detection in real-world field conditions.
- Li et al. (2024) fused multispectral and RGB data for canopy-scale classification but relied on regional labels rather than temporal, pixel-level disease mapping.
- Bleasdale & Whyatt (2025) and Turkoglu et al. (2019) demonstrated that NIR bands improve early scab detection, but applied CNNs independently to each image, without modeling temporal changes or segmentation.
- Patle et al. (2022) used LSTM-based time-series sensor data for mango disease prediction, achieving high accuracy; however, the absence of visual input limits spatial tracking of disease symptoms.
- Kabir et al. (2025) reviewed advanced greenhouse and smart farming techniques for apple production in controlled environments, emphasizing climate control and IoT-based monitoring. However, the study does not address image-based disease detection or pixel-level analysis

In contrast, our study addresses these gaps by combining multispectral time-series data with a U-Net + ConnvLSTM architecture, producing temporally informed, disease specific, pixel-level segmentation maps under challenging field conditions. This approach enables robust, temporally informed, pixel-level disease monitoring in commercial orchards and vineyards, addressing the limitations of prior RGB-based or image-level methods.

Table 1- Comparison of Related Studies

Study	*Input Data*	*Target Disease*	*Model/Method*	*Task*	*Performance*	*Setup & Limitation*
Bleasdale & Whyatt (2025)	RGB + NIR multispectral + RGB time series	Apple scab	MobileNetV2 & EfficientNetV2L CNN	Image-level classification	97.57% / 78.91% Accuracy (secondary vs multispectral data)	Close range imaging; image-level classification only, no pixel-wise segmentation
Verma & Dubey (2021)	RGB video	Rice brown spot, leaf blast	LSTM SRNN (NARX)	Image-level classification	99.99% Accuracy	Smartphone recorded videos; no pixel-level spatial localization
Gnanasaravanan et al. (2021)	RGB images	Various foliar diseases	BiLSTM	Image-level classification / BiLSTM	95.79% Accuracy	Close range imaging; Limited to binary classification, no pixel-level spatial localization
Farooqui et al. (2023)	RGB images	Apple scab, black rot, cedar apple rust, etc.	HG-ROA optimized LSTM	Image-level classification	99.12% Accuracy	Close range imaging; no pixel-level spatial localization
Madeira et al. (2024)	Satellite RGB + weather data	Vineyard leaf diseases (leaf blight, black rot, ESCA)	MobileNet V2	Image-level classification	99.375% Accuracy	Satellite imagery; coarse spatial resolution, no pixel-level spatial localization
Sundravadivelu et al. (2023)	RGB, + multispectral + thermal	General plant health (nutrient deficient)	Combined CNN	Image-level classification	99.20% Accuracy	Drone imagery; no disease specific segmentation, only general health classification
Ashwini & Sellam (2024)	RGB maize datasets	Maize leaf diseases (PLS, GLS, SR, NCLB, CR, MSV, FAW)	3D-CNN + BiLSTM	Image-level classification	99.12% Accuracy	Close range imaging; RGB only data, Image-level classification only, no pixel-level spatial localization
Bleasdale (2024)	RGB + NIR multispectral time series	Apple scab	MobileNetV2, InceptionResNetV2, EfficientNetV2L	Image-level classification	95.4% Accuracy	Close-range imaging; no full-scale orchard validation, no pixel-wise segmentation
Rous et al. (2023)	Field RGB + VIS NIR (8-band)	Apple scab	YOLOv5x with ANN/LDA preprocessing	Object detection	F1-score 0.70, mAP@0.5 = 0.73	Close-range imaging; sensitive to outdoor illumination and shadows, causing false positives
Zhang et al. (2025)	Close-range hyperspectral images (400-1000 nm, 76 bands)	Apple quarantine diseases (C. mali, M. fructicola, B. dothidea)	HSC-ResNet hyperspectral CNN with CBAM	Image-level classification	95.51% Accuracy	Laboratory conveyor setup; image-level classification only, no pixel-wise segmentation
Zhang et al. (2025)	Multispectral + chlorophyll fluorescence images	Verticillium Wilt in Eggplants	2D-CNN with multispectral CHIF image fusion	Image-level classification	99.37% Accuracy	Drone imagery, Greenhouse-only benchtop imaging, no in-field validation, no pixel-wise segmentation
Mydhili et al. (2025)	RGB	Apple dark stains due to internal feeding worm	DDA-ACV-LMWNN (LMWNN + LSIMC + TOA)	Image-level classification	99.60% Accuracy	Close-range imaging; Binary grading only, RGB data from controlled setup, no pixel-wise segmentation
Li et al. (2024)	RGB + vegetation indices	Apple canopy diseases and pests (Aphids, Alternaria, Mosaic, Brown spot, Gray spot)	AMMFNet	Image-level classification	92.92% Subset accuracy, 86.21% F1-score	Drone imagery; Regional outputs only, no pixel-wise segmentation

Table 1 (continue)- Comparison of Related Studies

Study	*Input Data*	*Target Disease*	*Model/Method*	*Task*	*Performance*	*Setup & Limitation*
Turkoglu et al. (2019)	High-res RGB	Apple diseases and pests (V. inaequalis, M. laxa, E. lanigerum, A. pomi)	MLP-CNN Ensemble	Image-level classification	99.20% Accuracy	High resolution, close range imaging; Image-level classification only, no pixel-level spatial localization
Patle et al. (2022)	Sensor time-series data	Mango powdery mildew, anthracnose, root rot	LSTM	Image-level classification / LSTM	96% Accuracy	Multi day sensor sequences; No visual data used, unsuitable for spatially resolved disease detection
This study	5 band multispectral time series images (Blue, Green, Red, Red Edge, NIR)	Apple scab + Grapes mildew	U Net + ConvLSTM (multispectral time series segmentation)	Pixel-level segmentation / ConvLSTM	Apple model: 82.05% mean IoU, 96.68% categorical accuracy; vineyard model: 79.67% mean IoU, 92.16% categorical accuracy.	Drone imagery; Requires 4 week sequential data of requested field before starting predictions

In contrast to previous studies, which predominantly rely on RGB imagery and image-level classification, the proposed approach performs pixel-level semantic segmentation of diseased regions using five-band multispectral drone imagery. The dataset includes Blue, Green, Red, Red Edge, and Near-Infrared (NIR) bands acquired with a DJI multispectral drone, providing substantially richer spectral information compared to conventional RGB-based approaches. The segmentation framework is built upon a U-Net backbone enhanced with LSTM units, enabling the exploitation of temporal dependencies derived from repeated drone flights. This design allows the model to generate spatially precise, disease-specific segmentation maps while also capturing the temporal progression of disease symptoms over time.

By integrating multispectral information, temporal modeling, and pixel-level segmentation within a unified framework, the proposed method offers a more comprehensive solution for field-scale disease monitoring. As summarized in Table 1, most existing studies are either limited to close-range or controlled imaging conditions, which are time-consuming and less suitable for operational deployment, or they provide only coarse-grained outputs such as image-level healthy/diseased classification without spatial localization or disease-specific mapping. Another common limitation observed in the literature is that models are typically trained and evaluated on data collected from the same field or environment, which may lead to overestimated performance and reduced generalization capability in real-world conditions.

3. Material and Methods

This section introduces the materials and methodological framework underlying the study.

3.1. Raw dataset

The raw datasets used in this study include multispectral drone images collected from apple orchards and vineyards in the agricultural areas of Tekirdag district, Turkey. The datasets include two disease types: apple scabs in the apple orchards and mildew in the vineyards. Data collection began on May 13, 2023, and continued until December 7, 2024. Although weekly data collection was initially planned, data was collected at irregular intervals due to weather conditions and operational limitations. As a result, the datasets cover a total time period of 29 weeks for apple orchards and 25 weeks for vineyards, providing sufficient temporal coverage for disease progression analysis. The apple orchard data covered a total of 29 weeks, with each imaging session consisting of five multispectral bands recorded on 25 fixed analysis grids in the monitored orchard blocks (29 weeks × 5 bands × 25 grids), resulting in a total of 3625 TIF samples. The vineyard data covered 25 weeks, consisting of five bands obtained on 45 analysis grids in the vineyard plots (25 weeks × 5 bands × 45 grids), corresponding to 5625 TIF samples. These multispectral images were captured as a time series, enabling temporal analysis of plant health across different stages. Time series data provides crucial information about the emergence of plant diseases, and designing a system based on this data allows for early diagnosis. The proposed model is designed to effectively capture both spatial and temporal features from multispectral time-series imagery. The model uses time series to extract features from changes in plants over time.

All vineyard samples contain mildew symptoms to reflect diseased conditions in the monitored parcels, whereas the apple images include both healthy and diseased trees together to better represent the natural variability of orchard conditions. Across both

sites, each grid corresponds to a fixed spatial region that was revisited in every flight, and pixel-level ground truth masks were created to distinguish diseased and healthy areas, enabling semantic segmentation rather than only image-level classification.

Images were captured using a DJI Phantom 4 Multispectral drone at an altitude of approximately 105 meters. Each image includes five spectral bands: Blue (450 ±16 nm), Green (560 ±16 nm), Red (650 ±16 nm), Red Edge (730 ±16 nm), and Near Infrared (840 ±26 nm). The raw images were saved in high resolution TIF format with a spatial resolution of 1600 × 1300 pixels. Figures 1 and 2 illustrate example images across different spectral bands.

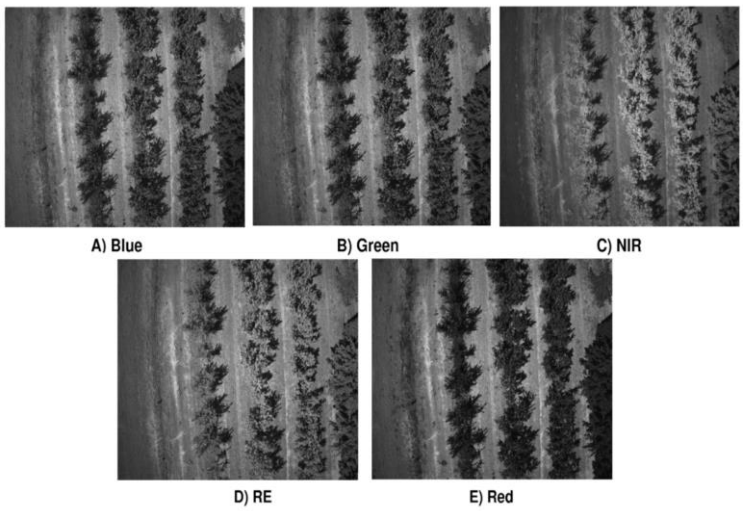

Figure 1- Multispectral drone images of apple orchards A) Blue Band, B) Green Band, C) NIR (Near Infrared) Band, D) RE (Red Edge) Band E) Red Band

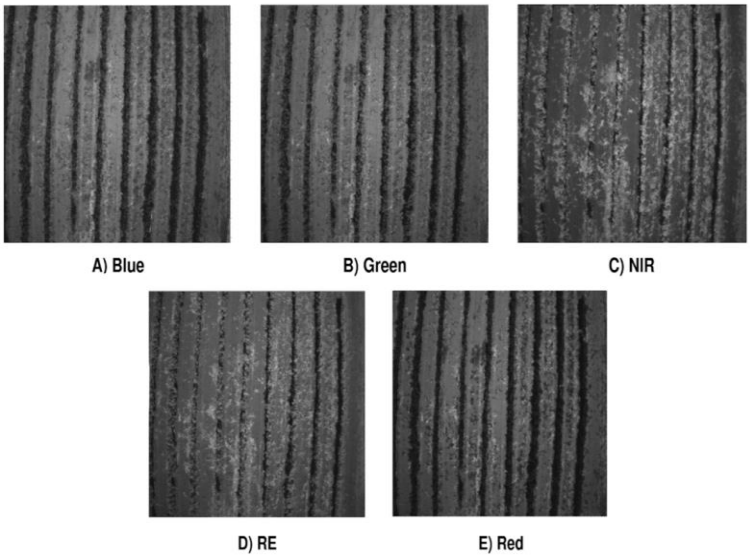

Figure 2- Multispectral drone images of vineyards A) Blue Band, B) Green Band, C) NIR (Near Infrared) Band, D) RE (Red Edge) Band E) Red Band

3.2. Data Preprocessing

The acquired multispectral images were preprocessed to ensure spatial and temporal consistency and to prepare the data for deep learning analysis. The main steps are summarized as follows:

- **Annotation (Ground Truth) Procedure:** Two agricultural engineers from Saricicek Agricultural Products Industry and Trade Inc., with domain expertise in apple and vineyard diseases, manually annotated diseased and healthy regions in the multispectral images. For selected grids, in-field observations were conducted to verify the correspondence between annotated regions and actual disease presence. The annotations were cross-validated between the two experts to minimize human error. Labels were assigned at the pixel level to enable semantic segmentation. Class imbalance between healthy

and diseased pixels was addressed during training using weighted loss functions. No automated or semi-automated labeling tools were used; all annotations were manually performed to ensure high-quality ground truth.

- **Temporal Consistency and Spatial Alignment:** To ensure temporal consistency, only continuous weeks of data were included in the analysis, and non-continuous observations were excluded. For each grid, sequences of four consecutive weeks were extracted from the 10-week observation period to construct temporally structured inputs for the ConvLSTM-based model. All images were acquired over fixed analysis grids using consistent flight paths. Furthermore, the multispectral imagery was georeferenced and orthorectified to ensure pixel-level spatial alignment across different time points. Visual validation using control points was performed to confirm spatial consistency across consecutive weeks. No missing temporal observations were present in the dataset, as only continuous weekly data were selected to ensure temporal consistency.

- **Imaging Parameters and Spatial Resolution:** Images were acquired using a DJI Phantom 4 Multispectral Drone equipped with five spectral bands (Blue, Green, Red, Red Edge, and Near-Infrared). The spatial resolution of the imagery was approximately 8 cm/pixel. Data collection covered 25 grid regions in apple orchards and 45 grid regions in vineyard plots, enabling accurate and spatially consistent pixel-level segmentation of disease-affected areas.

- **Normalization and Rescaling:** Pixel values for each spectral band were normalized to the range [0,1] to standardize input across all images. Images were resized to 160 × 200 pixels to ensure uniform input size while retaining sufficient spatial resolution for segmentation.

- **Data Augmentation:** To improve model generalization and mitigate the limited dataset size, data augmentation techniques were applied during training, including random rotations, horizontal and vertical flips. These transformations were applied consistently across all spectral channels to preserve the spatial and spectral integrity of the multispectral data.

- **Dataset Splitting:** The dataset was divided into training, validation, and testing sets in an 80%–10%–10% ratio. The test set was constructed from the final week to ensure temporally consistent and unbiased evaluation.

- **Temporal Sequence Construction:** For training, the data were organized into overlapping 4-week sequences (203 × 4 × 160 × 200 × 5), resulting in 203 spatio-temporal samples. This structure preserves both spatial and temporal information, enabling the model to learn disease progression dynamics effectively.

These preprocessing steps ensure the reliability of the ground truth annotations and the spatial-temporal consistency required for robust training and evaluation of the proposed ConvLSTM-U-Net model.

3.3. Performance metrics

To evaluate the segmentation and classification performance of the proposed model, the following metrics were used:

- **Intersection over Union (IoU):** Measures the overlap between predicted and ground-truth masks for pixel-level segmentation.

- **Categorical Accuracy:** Ratio of correctly classified pixels to total pixels.

These metrics allow quantification of both spatial and categorical accuracy of the proposed model. Although comparisons with alternative architectures such as standard U-Net or U-Net+LSTM could provide further insight, conducting such comparisons would require training separate models and is beyond the scope of this study. Future work may include these experiments. The current results focus on evaluating the performance of the proposed hybrid model and comparing it with relevant literature benchmarks.

3.3.1 Intersection over union metric (IoU)

The primary evaluation criterion used to assess the performance of the developed model is the Intersection over Union (IoU) metric. This metric, which quantifies the overlap between the predicted segmentation masks and the ground truth masks, is widely used in image segmentation object detection tasks. It measures the degree of overlap between the predicted bounding box or segmented area and the corresponding actual bounding box or actual segmented area. Mathematically, IoU is calculated as the ratio of the intersection area between two boxes or segmented areas to the area occupied by their merge. A higher IoU value indicates greater overlap between the predicted box and the actual box or segmented area. This means the trained model can make more accurate positioning predictions. This metric is particularly important in applications where precise object positioning directly impacts subsequent decision making processes, such as autonomous driving, medical image analysis, and video surveillance systems.

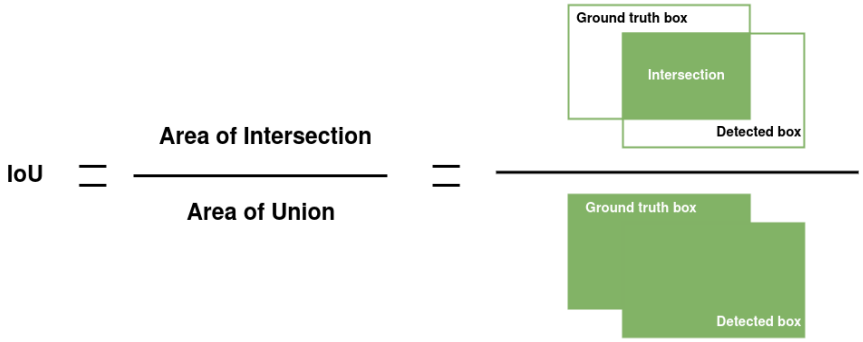

Figure 3- Intersection over Union Metric (IoU)

Figure 3 shows a visual representation of the IoU calculation process. In the upper part of the diagram, the overlapping region between the predicted bounding box and the actual box is shaded to show the intersection area. In the lower part of the diagram, the combination of both boxes is highlighted, representing the combined area used in the denominator of the IoU formula. IoU provides a detection accuracy measure that enables consistent evaluation and comparison of different detection models across diverse datasets and experimental setups by comparing the intersection/union ratio.

On the apple orchard test set, the model achieves a mean IoU of 82.05%, while the model reaches 79.67% on the vineyard mildew test set, indicating strong segmentation performance in both cases.

3.3.2. Categorical accuracy

Categorical accuracy is used alongside the Intersection over Union (IoU) metric to give a more complete view of performance. While IoU measures the spatial overlap between prediction and ground truth for segmentation, categorical accuracy reflects how often the class labels are predicted correctly. Categorical accuracy allows for the evaluation of classification performance at the pixel level.

Categorical Accuracy is a widely used evaluation metric for classification tasks, particularly in multi-class problems.

$$Categorical\ \ Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Observations} \tag{5}$$

In this formulation, True Positives represent the number of samples where the model predicted positive are actually positive and True Negatives represent the number of samples where the model predicted negative are actually negative.

To further optimize model accuracy, particularly in minimizing false positives, extensive experimentation was conducted by varying the temporal grouping of input data. Among the different configurations tested, the grouping of time series data into 4 week intervals yielded the lowest false positive rate, thus proving to be the most effective configuration. On the apple test set, the model achieved a categorical accuracy of 96.68%, and on the vineyard test set reached 92.16%, showing that most pixels are assigned to the correct class.

3.4. Deep learning architectures

This subsection introduces the deep learning-based architectures, U-Net, LSTM, and ConvLSTM, employed in this study and the proposed deep learning-based model.

3.4.1. U-Net architecture

U-Net, introduced by Ronneberger et al. (2015), is a convolutional neural network architecture originally designed for biomedical image segmentation. The encoder-decoder structure with skip connections enables efficient localization and context capture, making it highly suitable for pixel-based segmentation tasks even with limited labeled data. Figure 4 illustrates U-Net architecture.

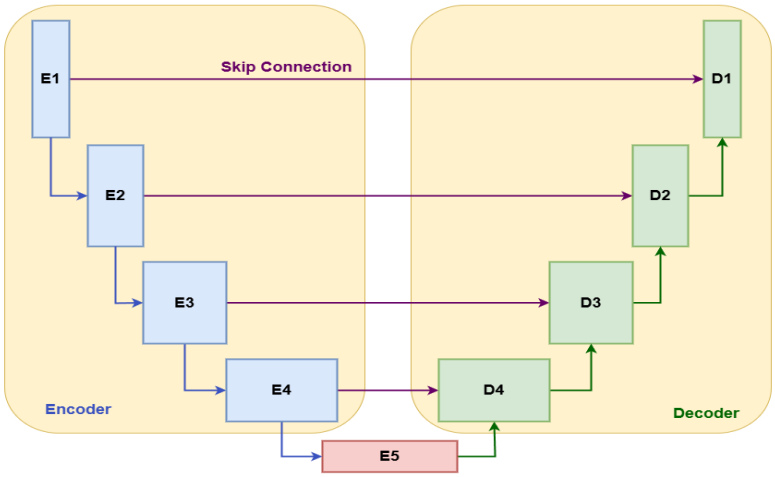

Figure 4- U-Net Architecture

3.4.2. Long Short-Term Memory (LSTM)

LSTM networks are a specialized type of recurrent neural network (RNN) capable of learning long-term dependencies in sequential data. The LSTM architecture consists of input, forget, and output gates, which regulate the flow of information, enabling the network to dynamically decide how much past information should be retained or discarded during the learning process. This structure has demonstrated superior performance in tasks where long-term dependencies are crucial, such as time series analysis, natural language processing, and video processing (Hochreiter & Schmidhuber 1997). Figure 5 illustrates LSTM structure.

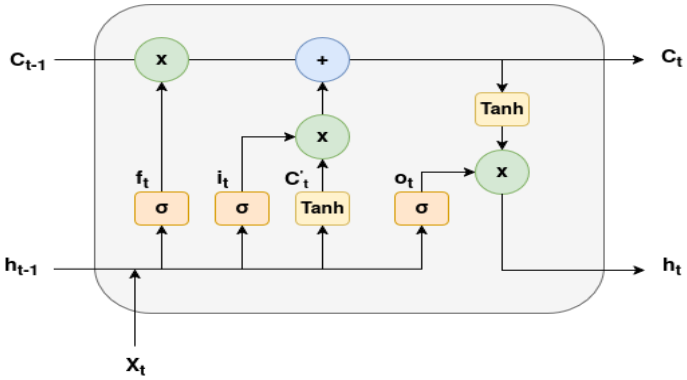

Figure 5- Internal structure of the LSTM

The internal structure of an LSTM cell consists of forget gate (f_t), input gate (i_t), candidated cell state (C'_t) , and the output gate (o_t) interacting with the current input (x_t), previous cell state (C_{t-1}) previous hidden state (h_{t-1}). Activation functions such as sigmoid (σ) and hyperbolic tangent ($Tanh$) are applied at different stages to control and transform the information flow, ensuring effective learning of both short-term and long-term patterns.

Step by step, the process works as follows:

1. **Forget gate(f_t)**: Takes the previous hidden state (h_{t-1}) and current input (x_t) and decides which information from the previous cell state (C_{t-1}) should be discarded.

2. **Input gate (i_t)**: Determines which new information will be added to the cell state, combining the sigmoid output with candidate values generated by the Tanh layer (C'_t).

3. **Cell state update (C_t)**: Combines the retained past information ($f_t . C_{t-1}$) with the new candidate information ($i_t . C'_t$).

4. **Output gate (o_t)**: Decides which part of the updated cell state will be exposed as the new hidden state (h_t), applying a $tanh$ transformation before gating the output.

Through this gated mechanism, LSTMs can capture contextual relationships across the temporal dimension of inputs, resulting in more accurate predictions. Furthermore, the flexibility of the gating mechanisms makes LSTMs more robust to noise in the data and effective in learning complex temporal dependencies. These advantages have established LSTMs as a preferred approach over conventional RNNs and simple feed-forward networks in numerous applications (Hochreiter & Schmidhuber 1997).

3.4.3. Convolutional Long Short-Term Memory (ConvLSTM)

ConvLSTM networks are an extension of traditional LSTMs. It is designed to handle spatiotemporal data by integrating convolutional structures into the recurrent architecture. Instead of using fully connected layers for state transitions, ConvLSTM replaces matrix multiplications with convolution operations and allows the network to preserve spatial structure while modeling temporal dependencies. This makes ConvLSTM particularly effective for applications involving sequential image or video data, where both spatial and temporal correlations are important (Shi et al. 2015). Figure 6 illustrates ConvLSTM structure.

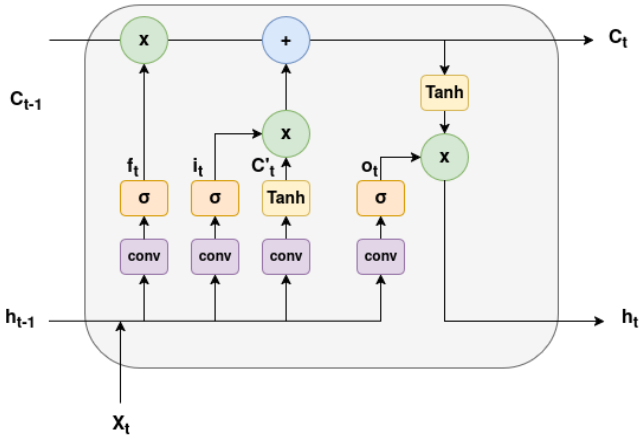

Figure 6- Internal structure of the Conv LSTM

The internal structure of a ConvLSTM cell represents an extension of the traditional LSTM architecture through the incorporation of convolutional operations. Similar to LSTM, the ConvLSTM cell retains the fundamental components of the forget gate (f_t), input gate (i_t), candidate cell state (C'_t), and output gate (o_t). However, in ConvLSTM, the matrix multiplications used in gate computations and state updates are replaced by convolution operations. This modification enables the network to preserve the spatial structure of the input data while simultaneously modeling temporal dependencies. Specifically, operations involving the current input (x_t), previous cell state (C_{t-1}) and the previous hidden state (h_{t-1}) are performed using convolutional filters instead of fully connected layers. As a result, ConvLSTM can process spatial and temporal features in an integrated manner, making it particularly effective for tasks involving sequential image or video data where maintaining spatial correlations is essential.

Step-by-step, the process operates as follows:

1. **Forget gate (f_t)**: Applies convolutional operations to the previous hidden state (h_{t-1}) and current input (x_t) to determine which information from the previous cell state (C_{t-1}) should be retained or discarded.
2. **Input gate (i_t)**: Uses convolutional operations on the current input and previous hidden state to determine the amount of new information to be added to the cell state. The sigmoid gate output is multiplied with the candidate cell state (C'_t) generated by the Tanh layer.
3. **Cell state update (C_t)**: Combines the retained past information ($f_t . C_{t-1}$) with the new candidate information ($i_t . C'_t$) to form the updated cell state.
4. **Output gate (o_t)**: Applies a $tanh$ transformation to the updated cell state, and multiplies it by the sigmoid output of the output gate to produce the new hidden state (h_t).

Through this convolution-based gating mechanism, ConvLSTM effectively captures both spatial and temporal contexts, leading to improved accuracy in tasks involving sequential visual data, such as video analysis and spatiotemporal prediction. Yin et al. (2023) successfully applied a hybrid U-Net + ConvLSTM model to monitor lake boundary changes with high accuracy, highlighting the potential of temporal modeling in geospatial tasks. Giannopoulos et al. (2022) further advanced this approach with a 4D U-Net that jointly learns spatio-spectro-temporal representations from Landsat-8 imagery, demonstrating improved performance in remote sensing classification tasks.

3.5. Proposed Model

This study introduces a novel deep learning architecture developed specifically for time series based disease detection and segmentation using multispectral drone imagery. The proposed model is designed to leverage the temporal and spectral richness of remote sensing data, capturing the dynamic progression of disease symptoms across multiple time points. In our case, to make a design-choose decision, we applied four different model variants on the Apple dataset by using different combinations of temporal length and activation function. We selected the most successful model from these variants as the base model. Table 2 shows the selected parameter combinations for each model variant, and the performance values of each variant are shown in Table 5 under Section 4.2. Comparative Analysis of Model Features.

Table 2- Ablation study on temporal length and activation functions

Configuration	***Temporal Length***	***Activation Functions of CNN***
Model Variant 1 (Proposed Model)	4 (weeks)	Swish
Model Variant 2	4 (weeks)	ReLU
Model Variant 3	3 (weeks)	Swish
Model Variant 4	3 (weeks)	ReLU

We utilize multispectral drone images consisting of five spectral bands, acquired at four distinct time intervals corresponding to weekly observations. Each input sample is therefore represented as a four-dimensional tensor of shape 4×160×200×5, where the first dimension corresponds to the temporal sequence length (i.e., four weeks), the spatial resolution of each image is 160×200 pixels, and the final dimension represents the five multispectral bands.

The overall structure of the model is based on an encoder decoder architecture enhanced with temporal modeling components and skip connections, as shown in Figure 7. The proposed network begins with a Convolutional Long Short-Term Memory (ConvLSTM) layer that acts as a spatio-temporal feature extractor. This layer is critical for modeling temporal dependencies between sequential image frames while simultaneously preserving spatial structure through convolutional operations. The ConvLSTM layer allows the model to learn how visual patterns representing plant diseases evolve over time, increasing its sensitivity to small temporal changes in multispectral data.

Following temporal encoding, the network undergoes an encoding (downscale) process consisting of three convolution layers. Each convolution block applies a convolution operation for feature extraction, followed by a downscaling step that reduces spatial dimensions. This structure allows for the learning of more abstract and meaningful features as the network deepens, while the resolution of the feature maps is gradually reduced. At the same time, thanks to the skip connections established between the encoder and decoder (upscale) layers, high resolution spatial information is preserved, enabling more accurate segmentation results.

The decoder path mirrors the encoder in structure and consists of three upscaling convolutional blocks. Each decoder block not only reconstructs the spatial resolution of the feature maps but also integrates contextual information received from the corresponding encoder block via the skip connections. This symmetric design facilitates the fusion of low-level spatial features with high-level abstract features, enabling precise localization of disease affected regions.

After the upscaling, two additional convolutional layers are applied as post processing modules to further refine the feature maps before the final segmentation. These layers ensure that the output segmentation mask is coherent and accurately delineated. Throughout the intermediate layers of the network, the Swish activation function (2) is employed to introduce non-linearity and enhance model expressiveness. The Swish function, proposed as an improvement over traditional ReLU based activations, is defined mathematically in the following equation.

$$Sigmoid\ (x) = \frac{1}{1+e^{-x}} \tag{1}$$

$$Swish\ (x) = x * Sigmoid(\beta x),\ \beta: scalable\ and\ trainable\ parameter \tag{2}$$

The multiplicative interaction between the input x and its sigmoid activation facilitates smoother gradient flow and contributes to more effective learning, particularly in deep architectures. For the final output layer, the Softmax activation function (3) is utilized to generate a pixel-wise probability distribution over the segmentation classes, ensuring that the model output is interpretable as class probabilities.

$$Softmax(z_i) = (e^{z_i})/(\sum_{j=i}^{k} e^{z_j}) \tag{3}$$

In this formulation, z_i denotes the raw score (logit) for class i, k represents the total number of classes, and j is the index used to sum over all class scores. The exponential operation transforms the logits into positive values, while the normalization term

ensures that the resulting probabilities for each pixel lie within the range [0,1] and sum to one across all classes, enabling meaningful multi class segmentation.

The algorithm for the proposed model is presented in Algorithm 1, which outlines the complete forward pass of the architecture. The process begins with temporal feature extraction using ConvLSTM layers to capture dynamic dependencies across sequential input frames. This is achieved by following an encoder path that reduces hierarchical sampling through a bottleneck block to deepen incremental convolutions and feature abstraction. The decoder symmetrically reconstructs spatial resolution via transpose convolutions, leveraging skip connections from the encoder to preserve localized spatial features. Finally a softmax activated output layer produces the segmentation mask, enabling dense pixel wise classification for multi class segmentation tasks.

Algorithm 1: Proposed Model

Input: **X**: A sequence of input frames of size **T×H×W×C** (**Time**, **Height**, **Width**, **Channels**)
Output: Segmented output mask of size **H×W×C**
BEGIN
// f: number of filters, k: kernel size, s: stride, p: same padding
// **Step 1**: **Temporal feature extraction and initial spatial encoding**
// f_{LSTM}: 20, k_{LSTM}: 3x3, s_{LSTM}: 1, X: 4 x 160 x 200 x 5
X_{LSTM} = ConvLSTM2D(X, f_{LSTM}, k_{LSTM}, s_{LSTM}, p)
X_{LSTM}= BatchNormalization(X_{LSTM})
X_{LSTM} = Swish(X_{LSTM})
// f_S: 32, k_S: 3x3, s_S: 1
X_S = Conv2D(X_{LSTM}, f_S, k_S, s_S, p)
X_S= BatchNormalization(X_S)
X_S = Swish(X_S)
// **Step 2**: **Encoder - Downscaling with strided convolutions**
FOR i = 1 to 3
  // **Conv2D** with:
  //$f_1 = 64, f_2 = 128, f_3 = 256$, k_i: 3x3, s_i: 2, $X_0 = X_S$
  X_i = Conv2D(X_{i-1}, f_i, k_i, s_i, p)
  X_i= BatchNormalization(X_i)
  X_i= Swish(X_i)
  E_i= X_i // Store for skip connection
END FOR
// **Step 3**: **Bottleneck convolution block**
// First conv in bottleneck: $f_{b1} = 256$, $k_{b1} = 3x3$, $s_{b1} = 1$
X_{b1} = Conv2D(X_3, $f_{b1} = 256$, $k_{b1} = 3x3$, $s_{b1} = 1$, p)
X_{b1} = BatchNormalization(X_{b1})
X_{b1} = Swish(X_{b1})
// Second conv in bottleneck: $f_{b2} = 256$, $k_{b2} = 3x3$, $s_{b2} = 1$
X_{b2} = Conv2D(X_{b1}, f_{b2}, k_{b2}, s_{b2}, p)
X_{b2} = BatchNormalization(X_{b2})
X_{b2} = Swish(X_{b2})
// **Step 4**: **Decoder - Upscaling** with **transpose convolutions**
FOR j = 3 to 1
  // **Conv2DTranspose**: $f_1 = 64, f_2 = 128, f_3 = 256$, $k_j = 3x3$, s_j: 2, $X_4 = X_{b2}$
  X_j = Conv2DTranspose(X_{j+1}, f_j, k_j, s_j, p)
  X_j = Concatenate(X_j, E_j) // Skip connection from encoder
  X_j = Conv2D(X_j, f_j, k_j, s_j, p)
  X_j = BatchNormalization(X_j)
  X_j= Swish(X_j)
END FOR
// **Step 5**: Output refinement and final prediction
// f_f: 32, k_f: 3x3, s_f: 1, $X_{prev} = X_1$
X_{out} = Conv2D(X_{prev}, f_f, k_f, s_f, p)
X_{out} = BatchNormalization(X_{out})
X_{out}= Swish(X_{out})
// f_{out}: 2 (equal to number of classes), k_{out}: 3x3, s_{out}: 1
Y = Conv2D(X_{out}, f_{out}, k_{out}, s_{out}, p)
Y_{output} = $Softmax(Y)$ // Final activation for multi-class segmentation
END

The detailed architectural configuration of the proposed model, including the number of filters, kernel sizes, and output dimensions at each stage, is presented in Table 3.

Table 3- Detailed architecture configuration of the proposed model

Stage	***Layer/ Block***	***Filters***	***Kernel Size***	***Stride***	***Activation***	***Output Size***
Input	Input Sequence	5 Channels	-	-	-	4 x 160 x 200 x 5
Temporal Feature Extraction	Convolutional LSTM 2D + Convolution Block	20 + 32	3x3	1	Swish	160 x 200 x 32
Encoder-1	Downsampling Convolution Block	64	3x3	2	Swish	80 x 100 x 64
Encoder-2	Downsampling Convolution Block	128	3x3	2	Swish	40 x 50 x 128
Encoder-3	Downsampling Convolution Block	256	3x3	2	Swish	20 x 25 x 256
Bottleneck	Convolution Block	256	3x3	1	Swish	20 x 25 x 256
Bottleneck	Convolution Block	256	3x3	1	Swish	20 x 25 x 256
Decoder-1	Transposed Convolution + Skip Connection + Convolution Block	128	3x3	2	Swish	40 x 50 x 256*****
Decoder-2	Transposed Convolution + Skip Connection + Convolution Block	64	3x3	2	Swish	80 x 100 x 128*****
Decoder-3	Transposed Convolution + Skip Connection + Convolution Block	32	3x3	2	Swish	160 x 200 x 64*****
Output Refinement	Convolution Block	2	3x3	1	Swish	160 x 200 x 32
Output	Final Convolution	2	3x3	1	Softmax	160 x 200 x 2

***:** After skip connections, the number of channels increases due to concatenation.

The step-by-step execution of the proposed model is formally described in Algorithm 1, and visually illustrated through the flowchart in Figure 7. The overall structural architecture of the model is presented in Figure 8, while the detailed layer-wise configuration is summarized in Table 3. The model begins by extracting temporal features through a ConvLSTM2D block, followed by downsampling via strided convolutions in the encoder. Intermediate features are preserved through skip connections and further processed in a bottleneck module. The decoder then reconstructs spatial resolution using transpose convolutions and integrates encoder features to enhance spatial fidelity. Finally, the output layer produces the segmented mask via softmax activation. Together, the algorithm and flowchart comprehensively illustrate the model's temporal to spatial processing pipeline.

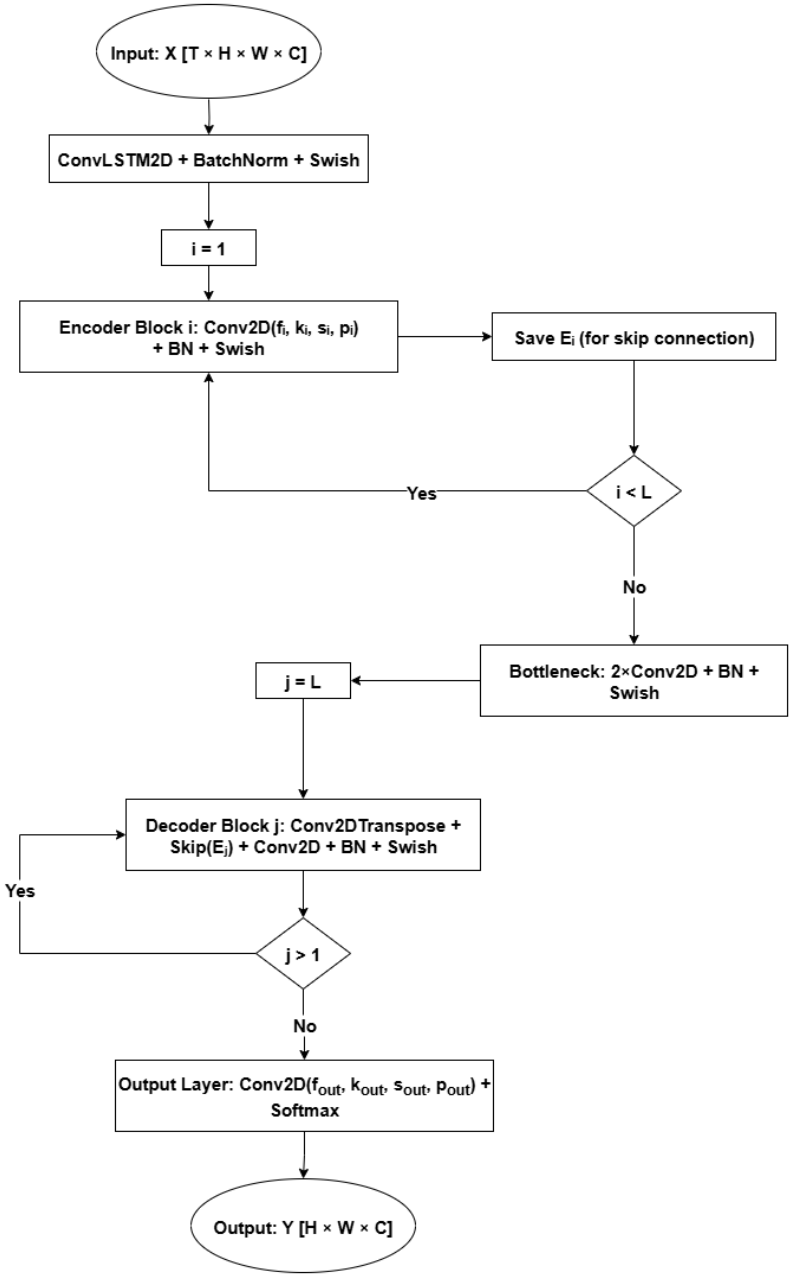

Figure 7- Flowchart of the proposed model

To optimize model training, we employ the Categorical Crossentropy loss function (4), which is well suited for multi class segmentation problems where the ground truth is encoded in a one-hot format.

$$Categorical\ Crossentropy\ (y, \hat{y}) = -\sum_{i=1}^{N} y_i . log(\hat{y}_i) \quad (4)$$

In this formulation, y_i denotes the ground truth label for class i (with a value of 1 for the correct class and 0 otherwise), $\hat{y}_i$ represents the predicted probability for class i, and N is the total number of classes. The loss is computed by taking the negative sum of the ground truth values multiplied by the logarithm of the corresponding predicted probabilities, penalizing deviations from the correct class and thereby guiding the model to improve its classification accuracy across all classes.

As illustrated in Figure 8, the proposed model integrates ConvLSTM layers with a U-Net based encoder-decoder framework to leverage both temporal and spatial modeling capabilities. The ConvLSTM component captures temporal dependencies and dynamic patterns within sequential input data, which is essential in applications where temporal coherence significantly influences prediction accuracy. The encoder-decoder structure performs hierarchical feature extraction through a series of convolutional downscale operations. Here, sequential convolution layers are used to gradually reduce spatial resolution while increasing the depth of feature representations. In the decoding phase, upscale is achieved using inverse convolution layers that reconstruct spatial resolution and allow the network to learn optimal sampling weights instead of relying on fixed interpolation

methods. Skip connections are used between the encoder and decoder layers. This improves localization and boundary accuracy by preserving high-resolution spatial information that might be lost during downscale. This hybrid design combines the temporal modeling capabilities of ConvLSTM with the spatial accuracy and structural robustness of an encoder-decoder architecture enhanced by skip connections, resulting in improved feature representation, better generalization, and more accurate output predictions across a variety of complex spatiotemporal tasks.

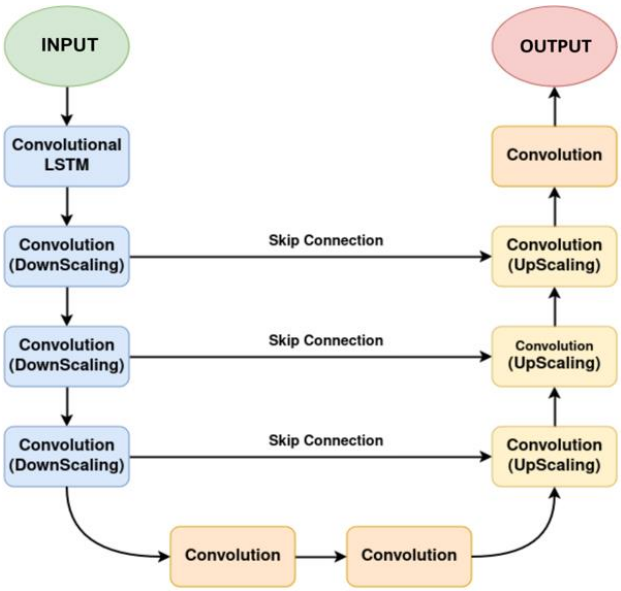

Figure 8- General structure of the proposed model

The proposed hybrid model combines the temporal modeling capabilities of ConvLSTM with the spatial precision of an encoder-decoder structure enhanced by skip connections and Swish activations as defined in (2). This design enables the model to effectively learn from both the spectral and temporal dimensions of multispectral drone imagery, providing robust and accurate segmentation of plant diseases over time.

In contrast to conventional U-Net + ConvLSTM implementations, which are typically designed for 3-channel RGB inputs and limited temporal configurations, the proposed model introduces two key extensions. First, the network is adapted to process 5-channel multispectral data (Blue, Green, Red, Red Edge, and NIR), enabling the extraction of richer spectral features that are particularly informative for vegetation analysis. Second, the ConvLSTM component is explicitly structured to model temporal dependencies across sequences of four consecutive weekly observations, allowing the network to capture the progression and evolution of disease symptoms over time. These enhancements enable the model to jointly leverage spectral and temporal information more effectively, thereby improving its representation capacity compared to standard approaches.

4. Results and Discussion

This section introduces the performance evaluation of the proposed model. In Subsection 4.1. Evaluation of Experimental Results, the effectiveness and practical applicability of the model are presented through both quantitative analysis and qualitative visual inspection of segmentation outputs, including training and validation performance trends and live test results. Subsection 4.2, Comparative Analysis of Model Features, provides a detailed comparison between the proposed architecture and existing models in the literature. Subsection 4.3. Discussion, further clarifies the model behavior and evaluation process. In particular, this section addresses the limitations related to dataset size and notes that the generalization capability of the model to other fields or crops may be restricted due to the limited dataset.

4.1. Evaluation of experimental results

This section provides a detailed evaluation of the experimental results of our proposed model. Figures 9 and 10 present the mean IoU scores across 50 training epochs for the model trained on the apple-dataset, and the model trained on the vineyard-dataset, respectively, with 4-week grouped time series data. In Figure 9, the validation curve initially remains relatively flat, then exhibits a sharp increase after approximately the 20th epoch, ultimately stabilizing at around 82.05%, which corresponds to the final performance achieved on the test dataset. This pattern indicates a delayed but effective learning phase, suggesting that the model requires several epochs to generalize from the training data. For vineyards, as shown in Figure 10, an IoU score of 79% was obtained after the first 14 epochs. After 50 epochs, an IoU score of 79.67% was achieved. The plateaus observed in Figures 9 and 10 confirm that the model has reached a stable level of segmentation performance and are suitable for practical use.

It is important to note that the achieved IoU values are influenced by inherent data acquisition constraints. Specifically, aerial imagery captured at an operational altitude of approximately 105 m limits the spatial resolution at the individual plant level, thereby reducing the availability of fine-grained structural details. In addition, the use of five-channel multispectral data restricts the spectral richness per plant. These limitations, together with factors such as occlusions and shadowing, impose a practical upper bound on the achievable segmentation accuracy, which explains why IoU values do not approach 90% under complex field conditions.

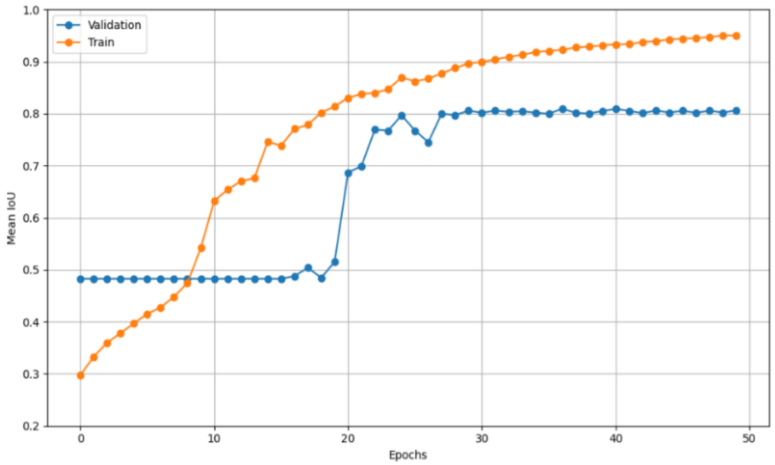

Figure 9- Training performance of the proposed model on the apple dataset

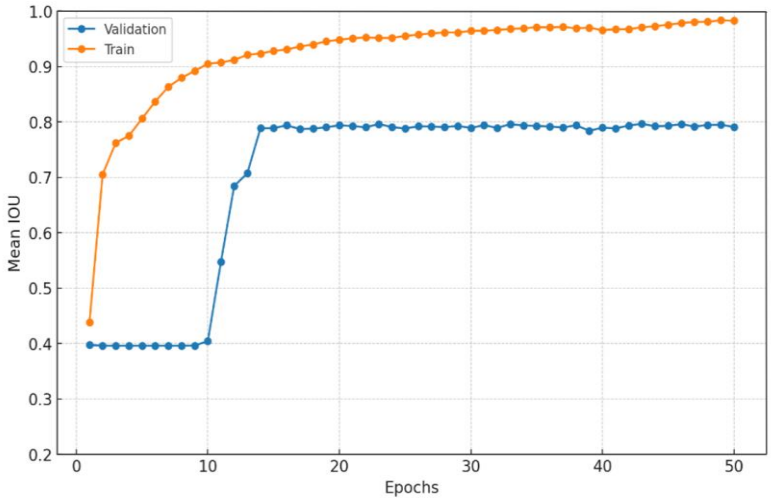

Figure 10- Training performance of the proposed model on the vineyard dataset

Figure 11 provides representative segmentation outputs generated by the proposed model trained on the apple dataset using 4-week grouped time series data. Figure 12 provides representative segmentation outputs generated by the proposed model trained on the vineyard dataset using 4-week grouped time series data. These examples showcase the alignment between predicted masks and actual disease regions, and further emphasize the model's practical applicability in real-world disease monitoring tasks.

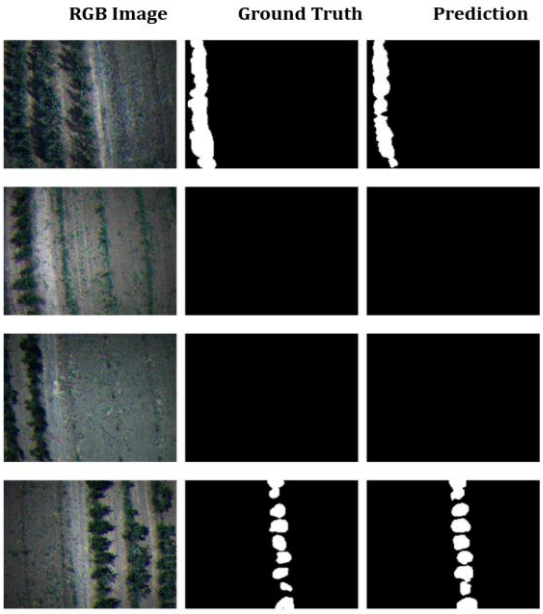

Figure 11- Comparison of the predictions of the proposed model with ground truth on the apple test data

In Figure 11, the proposed hybrid model gives segmentation masks that are very close to the ground truth. The first column shows image, the second column the expert-labeled diseased areas, and the third column the model's output. The visual agreement between the masks indicates that the model can detect and outline diseased regions reliably.

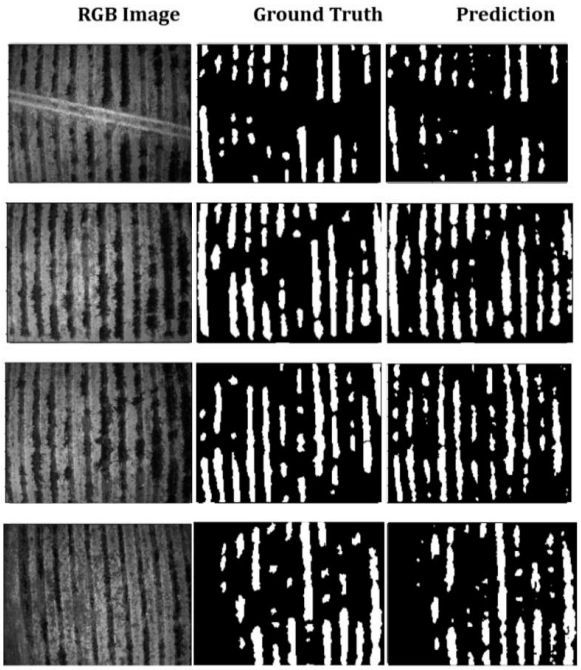

Figure 12- Comparison of the predictions of the proposed model with ground truth on the vineyard test data

In Figure 12, images, ground truth mildew masks, and model predictions are again shown together to illustrate how the model segments mildew-infected grape canopies.

In Figures 13 and 14, the live test results of the proposed model are presented, demonstrating the transfer of the segmentation capabilities shown in Figures 11 and 12 to real-world field conditions. The orange-marked areas indicate regions identified by the model as disease affected, overlaid onto the original drone imagery. These results show that the model can maintain high detection performance not only on prepared test datasets but also in operational environments under varying illumination and environmental conditions. The ability to localize diseased plants in real time highlights the model's potential for practical implementation in large scale agricultural disease monitoring.

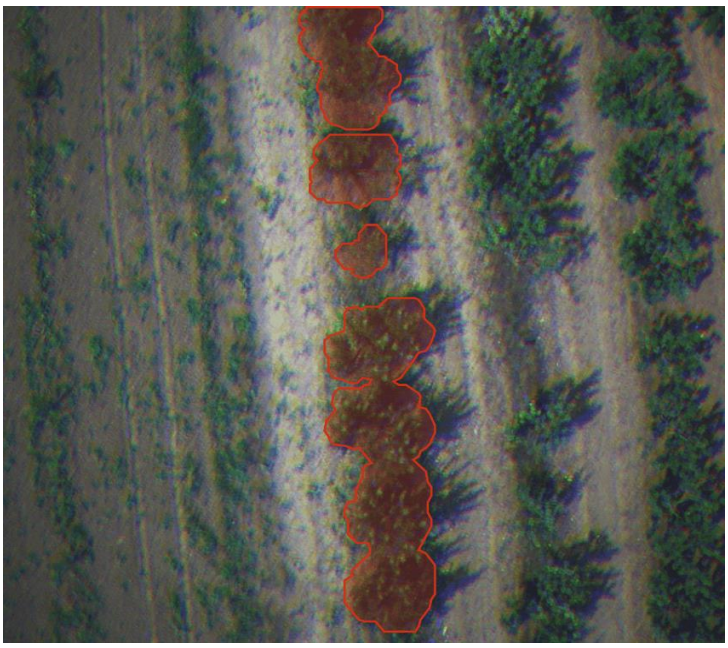

Figure 13- Live detection results of the proposed model on apple orchard images (detected apple scab regions outlined in red)

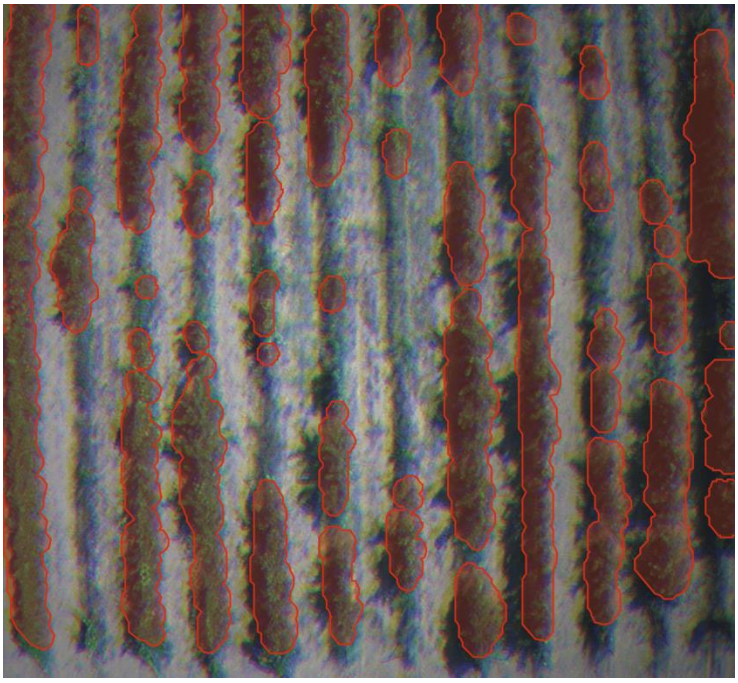

Figure 14- Live detection results of the proposed model on vineyard images (detected mildew infected regions outlined in red).

4.2. Comparative analysis of model features

To contextualize the architectural and functional advancements of our proposed model, a comparative analysis was conducted against several recent and representative deep learning approaches in the literature. As summarized in Table 4, the comparison focuses on four critical capabilities: usage of RGB image data, inclusion of multispectral image data, incorporation of temporal (time series) data, and suitability for segmentation tasks. The vast majority of current models rely on RGB images and address classification tasks without considering temporal dependencies or spectral diversity. In contrast, our proposed model uniquely integrates multispectral input, time-dependent sequence modeling via ConvLSTM, and pixel-level segmentation through a UNet-inspired encoder-decoder architecture. These features of the proposed method offer a robust and generalizable solution for complex spatio-temporal tasks where both spectral richness and fine-grained temporal variations play a crucial role in accurate and interpretable predictions.

Table 4- Comparison of the Proposed Model with Existing Approaches

Study/Model	*RGB Image Data*	*Multispectral Image Data*	*Temporal Data (Time Series)*	*Segmentation Task*
Proposed Model	NO	**YES**	**YES**	**YES**
Combined CNN (Sundravadivelu et al. 2023)	**YES**	**YES**	NO	NO (Classification)
LSTM SRNN (NARX) (Verma & Dubey 2021)	**YES**	NO	**YES**	NO (Classification)
HG-ROA optimized LSTM (Farooqui et al. 2023)	**YES**	NO	**YES**	NO (Classification)
3D-CNN + BiLSTM (Ashwini & Sellam 2024)	**YES**	NO	**YES**	NO (Classification)
BiLSTM (Gnanasaravanan et al. 2021)	**YES**	NO	**YES**	NO (Classification)
MobileNet V2 (Madeira et al. 2024)	**YES**	NO	NO	NO (Classification)
MLP-CNN Ensemble (Turkoglu et al. 2019)	**YES**	NO	NO	NO (Classification)
LSTM (Patle et al. 2022)	NO	NO	**YES**	NO (Classification)
MobileNetV2 + EfficientNetV2L CNN (Bleasdale & Whyatt 2025)	**YES**	**YES**	NO	NO (Classification)
MobileNetV2 + InceptionResNetV2 + EfficientNetV2L CNN (Bleasdale 2024)	**YES**	**YES**	NO	NO (Classification)
YOLOv5x (Rous et al. 2023)	**YES**	**YES**	NO	NO (Object Detection)
HSC-ResNet (Zhang et al. 2025)	NO	**YES**	NO	NO (Classification)
2D-CNN Fusion Model (Zhang et al. 2025)	NO	**YES**	NO	NO (Classification)
AMMFNet (Li et al. 2024)	**YES**	**YES**	NO	NO (Classification)
Deep Learning Saliency Map (Mydhili et al. 2025)	**YES**	NO	NO	**YES**

As shown in Table 4, our proposed model distinguishes itself as the only approach that simultaneously utilizes RGB and multispectral image data, incorporates temporal dynamics through time series modeling, and performs pixel level segmentation rather than purely classification. This comprehensive capability addresses a critical gap in the literature, where most models focus on a limited subset of these features. For instance, the Combined CNN model (Sundravadivelu et al. 2023) supports both RGB and multispectral input, but lacks temporal modeling and does not perform segmentation, its use case is confined to static frame classification. Similarly, methods like LSTM SRNN (Verma & Dubey 2021), HG ROA optimized LSTM (Farooqui et al. 2023) and 3D-CNN + BiLSTM (Ashwini & Sellam 2024) capture temporal dependencies to some extent, yet fail to leverage multispectral richness and are not equipped for pixel wise spatial delineation. This partial integration limits their effectiveness in tasks that require simultaneous spectral, spatial, and temporal reasoning, such as monitoring plant health progression or detecting plant disease spread over time in high resolution imagery.

Most of the models examined, including BiLSTM (Gnanasaravanan et al. 2021), MobileNet V2(Madeira et al. 2024), MLP CNN Ensemble (Turkoglu et al. 2019) and LSTM (Patle et al. 2022) are designed to focus primarily on RGB-based static image classification, lacking both temporal context and multispectral analysis. This significantly limits the usefulness of these methods in applications requiring precise spatial positioning. In contrast, our model's hybrid architecture, which integrates ConvLSTM modules within a UNet based encoder decoder approach, enables the network to simultaneously learn temporal evolution, spectral relationships and spatial hierarchies. The proposed model effectively combines spectral, temporal, and spatial information to enable more detailed decision-making, stronger generalization over time, and more accurate segmentation of complex patterns within the input data. This integrated design not only improves predictive performance but also makes the model well suited for applications that require high segmentation accuracy, temporal consistency, and sensitivity to multispectral inputs. In addition to comparisons with existing methods, an ablation study was conducted to evaluate the impact of temporal sequence length and activation functions on the performance of the proposed model. The evaluated configurations were derived from controlled variations of the baseline architecture.

In addition to comparisons with existing methods, an ablation study was conducted to evaluate the impact of temporal sequence length and activation functions on the performance of the proposed model. The evaluated configurations were derived from controlled variations of the baseline architecture.

Table 5- Ablation study results for different temporal sequence lengths and activation functions

Configuration	***Temporal Length***	***Activation Functions of CNN***	***Mean IoU (Apple)***
Model Variant 1 (Proposed Model)	4 (weeks)	Swish	82%
Model Variant 2	4 (weeks)	ReLU	80%
Model Variant 3	3 (weeks)	Swish	81%
Model Variant 4	3 (weeks)	ReLU	79%

As shown in Table 5, the base configuration of the proposed model, which utilizes a 4-week temporal sequence and the Swish activation function, achieves the highest segmentation performance. Variants with a reduced temporal length (3 weeks) consistently exhibit lower performance, indicating that a longer temporal context improves the model's ability to capture disease progression patterns.

Furthermore, replacing the Swish activation function with ReLU leads to a consistent decrease in performance across both temporal configurations, highlighting the importance of activation function selection for effective feature representation and learning dynamics.

Overall, these findings confirm that both temporal sequence length and activation function choice are critical design factors for robust spatio-temporal segmentation. The ablation study was conducted on the apple dataset to determine the optimal model configuration, which was subsequently applied to the vineyard dataset to ensure consistency in evaluation.

4.3. Discussion

This study demonstrates that integrating multispectral imagery with temporal modeling through a hybrid U-Net + ConvLSTM architecture provides a robust framework for early-stage plant disease detection under real-world field conditions. The achieved IoU values of 82.05% for apple orchards and 79.67% for vineyards indicate that the proposed model can produce reliable pixel-level segmentation despite challenges such as varying illumination, occlusions, and reduced spatial resolution at higher flight altitudes. Unlike conventional approaches that rely on RGB-based image-level classification, the proposed method jointly exploits spectral richness and temporal progression, enabling the detection of subtle physiological changes before they become visually apparent. These findings highlight the importance of integrating spatial, spectral, and temporal information within a unified deep learning framework for precision agriculture applications.

From a modeling perspective, the integration of ConvLSTM into the U-Net backbone plays a critical role in capturing temporal dependencies across sequential observations. The ablation study further demonstrates that both temporal sequence length and activation function selection significantly influence performance, with a 4-week temporal configuration and the Swish activation function yielding the best results. This suggests that disease progression patterns require sufficient temporal context to be effectively modeled. In addition, the encoder–decoder structure with skip connections preserves both high-level semantic features and low-level spatial details, contributing to accurate boundary delineation in segmentation outputs. However, the lack of direct comparisons with alternative architectures, such as standard U-Net or purely CNN-based models, limits the ability to isolate and quantify the contribution of each architectural component. This limitation represents an important direction for future research.

From a data perspective, the use of drone-acquired multispectral time-series data provides a strong basis for capturing disease dynamics across different growth stages. The fixed grid-based acquisition strategy and consistent flight paths ensure spatial and temporal alignment, facilitating reliable model training. Nevertheless, the relatively limited dataset constrains the model's

generalization capability. The data were collected from specific regions in Tekirdağ and reflect particular environmental conditions, crop types, and disease characteristics, which may not fully represent broader agricultural variability. To mitigate this limitation, data augmentation techniques—such as random rotations and horizontal and vertical flips—were applied during training. These transformations were consistently implemented across all spectral channels to preserve spatial and spectral coherence, thereby enhancing model robustness to variations in orientation and spatial distribution. Despite these efforts, the limited diversity of training data may partially explain why IoU values do not exceed very high thresholds (e.g., >90%), as unseen field variability is not fully captured.

Despite these promising results, several limitations should be acknowledged. Since the dataset was collected from the same agricultural fields using a fixed grid structure, there is a potential risk of spatial bias and data leakage. In such cases, the model may inadvertently learn location-specific patterns rather than fully generalizable features. Although care was taken to separate training and testing samples to minimize direct overlap, spatial correlations within the same field may still persist.

Future work will focus on mitigating these limitations by adopting spatially disjoint training–testing strategies, such as field-wise or region-wise splits, and by validating the model on independent datasets collected from different fields, seasons, and environmental conditions. These steps will enable a more reliable assessment of the model's generalization capability and support the broader applicability of the proposed approach.

5. Conclusions

In this study, a new deep learning-based hybrid model was presented for the early detection of plant diseases, specifically apple scab disease in apple orchards and mildew disease in vineyards. The proposed deep learning-based architecture was developed by integrating LSTM layers with a CNN backbone within the customized U-Net framework. The proposed model architecture was specifically designed to leverage time series information, enabling robust tracking of crop health changes over time. Using time series data collected between May 2023 and December 2024 by a DJI Phantom 4 Multispectral drone providing five different spectral bands in agricultural areas of Tekirdag province, Turkey, and processing this data with the proposed model, disease-affected areas in apple orchards and vineyards were identified and segmented with high accuracy. The proposed model demonstrated a strong classification ability in accurately predicting class labels, achieving categorical accuracy rates of 96.68% and 92.16% in the apple and vineyard datasets, respectively. Experimental results show that the proposed model achieved an average Intersection Over Join (IoU) of 82.05% in the apple dataset and 79.67% in the vineyard dataset, demonstrating strong segmentation performance.

The findings highlight the effectiveness of the proposed model in monitoring plant health in precision agriculture when applied by combining spectral richness and temporal context in multispectral images. This approach not only supports early and reliable disease detection but also contributes to informed decision making for plant management over time.

Author Declarations

Authors' contributions: Concept: B.A., F.S.B., S.S., Design: B.A., A.K.T., F.S.B, S.S. Data collection: B.A., A.K.T., F.S.B., Analysis: B.A., A.K.T., F.S.B., S.S., Writing: B.A, S.S.

Conflict of interest/Competing interests: The author(s) declare no conflict of interest.

Ethics approval/declarations: No need/ Not applicable.

Funding: This research received no external funding.

Consent to participate: No need/ Not applicable.

Consent for publication: No need/ Not applicable.

Data availability: Data are available on request due to privacy or other restrictions.

References

Ashwini C & Sellam V (2024). An Optimal Model for Identification and Classification of Corn Leaf Disease Using Hybrid 3D-CNN and LSTM. Biomedical Signal Processing and Control, 92:106089. https://doi.org/10.1016/j.bspc.2024.106089

Avotins A, Kviesis K, Bicans J, Alsina I, & Dubova L (2020). Experimental analysis of IoT based camera SI-NDVI values for tomato plant health monitoring application. Agronomy Research, 18(S2): 1138-1146. https://doi.org/10.15159/AR.20.087

Bebber D P, Castillo Á D & Gurr S J (2020). Modelling coffee leaf rust risk in Colombia with climate reanalysis data. Philosophical Transactions of the Royal Society B: Biological Sciences, 375(1810): 20190578. https://doi.org/10.1098/rstb.2019.0578

Bleasdale A J (2024). The early detection of apple scab using multispectral imagery under natural illumination conditions (Doctoral thesis). Lancaster Environment Centre, Lancaster University, Lancaster, UK.

Bleasdale A J & Whyatt J D (2025). Classifying early apple scab infections in multispectral imagery using convolutional neural networks. Artificial Intelligence in Agriculture, 15: 39–51. https://doi.org/10.1016/j.aiia.2024.10.001

Cao X, Zheng S, Zhang J, Chen W & Du G (2025). A hybrid CNN-Bi-LSTM model with feature fusion for accurate epilepsy seizure detection. BMC Medical Informatics and Decision Making, 25: 6. https://doi.org/10.1186/s12911-024-02845-0

Chen X, Zhou G, Chen A, Yi J, Zhang W & Hu Y (2020). Identification of tomato leaf diseases based on combination of ABCK-BWTR and B-ARNet. Computers and Electronics in Agriculture, 178, 105730. https://doi.org/10.1016/j.compag.2020.105730

De Silva M & Brown D (2022). Plant disease detection using multispectral imaging. In International Advanced Computing Conference (pp. 290-308). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35641-4_24

Farooqui N A, Mishra A K & Mehra R (2023). Concatenated Deep Features with Modified LSTM for Enhanced Crop Disease Classification. *International Journal of Intelligent Robotics and Applications*, 7: 510–534. https://doi.org/10.1007/s41315-022-00258-8

Giannopoulos M, Tsagkatakis G & Tsakalides P (2022). 4D U-Nets for Multi-Temporal Remote Sensing Data Classification. Remote Sensing, 14(3): 634. https://doi.org/10.3390/rs14030634

Gnanasaravanan S, Tharani B & Sahu M (2021). Long Short-Term Memory Recurrent Neural Networks for Plant Disease Identification. International Journal of Computer Communication and Information, 3(2): 51–62. https://doi.org/10.34256/ijcci2125

Hochreiter S & Schmidhuber J (1997). Long short-term memory. Neural Computation, 9(8): 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hua Y & Liu G (2024). Food Pesticide Residues Monitoring and Health Risk Assessment. Foods, 13(3): 474. https://doi.org/10.3390/foods13030474

Kabir M H, Tan C & Zhang Z (2025). Apple Production in Controlled Environments: Greenhouse Techniques and Best Practices. In: Zhang, Z., Zhu, D., Zuo, C., Han, B., Liu, P., Wang, Z. (eds) Apple Production Technologies: From Laboratory to Practical Applications. Smart Agriculture, 12. Springer, Singapore. https://doi.org/10.1007/978-981-96-5747-6_1

Li H, Tan B, Sun L, Liu H, Zhang H & Liu B (2024). Multi-source image fusion based regional classification method for apple diseases and pests. Applied Sciences, 14(17): 7695. https://doi.org/10.3390/app14177695

Madeira M, Porfírio R P, Santos P A & Madeira R N (2024). AI-powered Solution for Plant Disease Detection in Viticulture. Procedia Computer Science, 238: 468–475. https://doi.org/10.1016/j.procs.2024.06.049

Mahlein A-K (2016). Plant disease detection by imaging sensors – Parallels and specific demands for precision agriculture and plant phenotyping. Plant Disease, 100(2): 241–251. https://doi.org/10.1094/PDIS-03-15-0340-FE

Mydhili S K, Senthilkumar K P, Buvaneswari B & Vijaya Lakshmi T R (2025). Detecting defective apples through advanced computer vision and Legendre Multi Wavelet Neural Networks. *Journal of Stored Products Research*, 114: 102702. https://doi.org/10.1016/j.jspr.2025.102702

Patle K S, Saini R, Kumar A & Palaparthy V S (2022). Field Evaluation of Smart Sensor System for Plant Disease Prediction Using LSTM Network. *IEEE Sensors Journal*, 22(4): 3715-3725. https://doi.org/10.1109/JSEN.2021.3139988

Ronneberger O, Fischer P & Brox T (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28

Rous R, Peller J, Polder G, Hageraats S, Ruigrok T & Blok P M (2023). Apple scab detection in orchards using deep learning on colour and multispectral images. Preprint submitted to Computers and Electronics in Agriculture. https://doi.org/10.48550/arXiv.2302.08818

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

Shahi T B, Xu C-Y, Neupane A & Guo W (2023). Recent Advances in Crop Disease Detection Using UAV and Deep Learning Techniques. Remote Sensing, 15(9): 2450. https://doi.org/10.3390/rs15092450

Shi X, Chen Z, Wang H, Yeung D Y, Wong W K & Woo W C (2015). Convolutional LSTM Network: a machine learning approach for precipitation nowcasting. In Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'15), MIT Press, Cambridge, MA, USA, 802–810. https://doi.org/10.48550/arXiv.1506.04214

Stamford J D, Vialet-Chabrand, S, Cameron I & Lawson T (2023). Development of an accurate low cost NDVI imaging system for assessing plant health. Plant Methods, 19: 9. https://doi.org/10.1186/s13007-023-00981-8

Sundravadivelu K, Senthilvel P G, Thirupurasundari D R, Rajesh Kumar K & Palani H K (2023). Automated Drone-Based Imaging Systems For Plant Health Monitoring Using Deep Learning Techniques. 2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS), Chennai, India, pp. 1-5. https://doi.org/10.1109/ICCEBS58601.2023.10449190

Turkoglu M, Hanbay D & Sengur A (2019). Multi-model LSTM-based Convolutional Neural Networks for Detection of Apple Diseases and Pests. Journal of Ambient Intelligence and Humanized Computing, 13(7): 3335–3345. https://doi.org/10.1007/s12652-019-01591-w

Verma T & Dubey S (2021). Prediction of Diseased Rice Plants Using Video Processing and LSTM-Simple Recurrent Neural Network. Multimedia Tools and Applications, 80: 29267–29298. https://doi.org/10.1007/s11042-021-10889-x

Yin L, Wang L, Li T, Lu S, Tian J, Yin Z, Li X & Zheng W (2023). U-Net-LSTM: Time Series-Enhanced Lake Boundary Prediction Model. Land, 12(10): 1859. https://doi.org/10.3390/land12101859MDPI+3MDPI

Zhang D, Luo S, Zhang J, Li M, Fan X, Chen X & Shen S (2025). Multispectral and chlorophyll fluorescence imaging fusion using 2D-CNN and transfer learning for cross-cultivar early detection of Verticillium wilt in eggplants. Agronomy, 15(8): 1799. https://doi.org/10.3390/agronomy15081799

Zhang H, Ye N, Gong J, Xue H, Wang P, Jiao B, Yin L & Qiao X (2025). Hyperspectral imaging-based deep learning method for detecting quarantine diseases in apples. Foods, 14(18): 3246. https://doi.org/10.3390/foods14183246

Copyright 2026 The Author(s). This is an open-access article published by Faculty of Agriculture, Ankara University under the terms of the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.