The 11th Asian-Australasian Conference on Precision Agriculture (ACPA 11)
October 14-16, 2025, Chiayi, Taiwan

PEST AND DISEASE IMAGE-TEXT IDENTIFICATION SYSTEM OF LEAFY VEGETABLES IN URBAN COMMUNITY FARMING

Chiao-Chi Hsu[1], Ting-Ting Li[2], Ya-Ching Yang[2], Po-Yi Wu[1], Shih-Fang Chen[1*]

[1] Department of Biomechatronics Engineering, National Taiwan University, Taiwan.
[2] Taoyuan District Agricultural Research and Extension Station, Ministry of Agriculture, Taiwan.

*Corresponding Author: sfchen@ntu.edu.tw

Abstract

Urban community farming has been integrated into education for sustainable food and agriculture. However, the participants are primarily students and novice farmers with limited background knowledge. Managing pests and diseases becomes challenging for these growers as diverse vegetable crops attract various pest and disease species, requiring accurate identification and treatment expertise. There is a need to develop timely identification services and guidance on control measures. In the field of pest and disease identification, considerable studies have been conducted on image-based methods, while showing limitations with poor-quality images. To address the identification challenges and explore the effectiveness of text in enhancing identification, this research developed a pest and disease identification system based on an image-text identification model, laying the foundation for online identification services. A multimodal dataset was established comprising 2,780 field vegetable images across 13 pest and disease categories, including six pest bodies, four feeding damage leaves, and three disease-infected leaves, along with the corresponding text of vegetable crops and environmental conditions. An image-text identification model utilizing fine-tuned YOLOv9 and RoBERTa-Chinese for visual feature extraction and text processing, respectively, was trained using the dataset. In the preliminary results, the proposed image-text identification model showed enhanced presence-based precision from 0.813 (image) to 0.929 (image-text), achieving comparable performance to a fine-tuned YOLOv9 (0.950), which demonstrates the potential of text-assisted image identification. However, when using image inputs, the model exhibited a high quantity of missed detections and misclassifications for three categories (striped flea beetle of 0.305; mustard leaf beetle adult and larva of 0.439 and 0.222, respectively), indicating room for improvement in image feature extraction and model architecture. The future works focus on the optimization of the model structure and expanding the dataset.

Keywords: Multimodal identification model, image-text identification, urban community farming, pest and disease identification.

INTRODUCTION

As a form of urban agriculture, community farming plays a key role in promoting sustainable urban food system. It provides shared spaces where residents collectively grow vegetable crops. Studies indicate that community farming improves food security in urban centers, also fosters positive social interactions, and increases public education about food system. In 2022, the Taiwanese government introduced legislation to institutionalize community farming initiatives across schools, communities, and civil organizations, aiming to enhance urban food security systems and advance public education on sustainable agriculture. However, it is a challenge for those novice growers to identify the pests and pathogens and develop targeted control strategies. The high vegetable crop diversity in community farms inevitably attracts a variety of pests and pathogens, complicating field management and requiring background knowledge. Accordingly, this research aims to develop a pest and disease identification system based on an image-text identification model by complementing visual analysis with descriptive text information. By combining multimodal inputs, this approach not only addresses the diagnostic challenges faced by community growers lacking formal training, but also improves the model's robustness and adaptability to diverse symptoms and image conditions.

MATERIALS AND METHODS

Pest and Disease Image Dataset

The image dataset comprises 2,780 images, encompassing 13 categories of pests and diseases, namely six pest bodies, four feeding damage leaves, and three disease-infected leaves (Fig. 1). To ensure diverse image sources and enhance model generalization, images were captured using both professional cameras and smartphones. The sampled vegetable crops included over 20 varieties such as cabbage, lettuce, eggplants, cucumber, tomato, and peas.

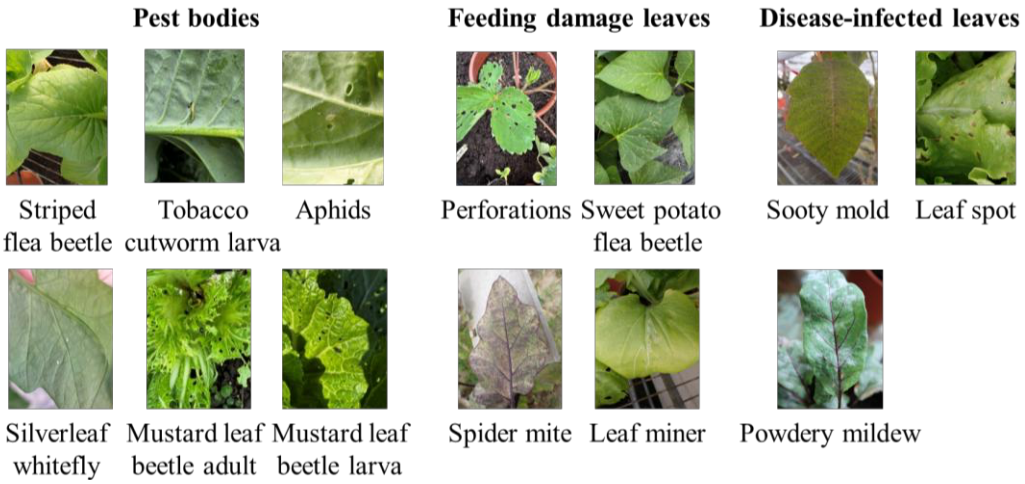

Fig. 1 Example images illustrating the three categories of the dataset.

Pest and Disease Text Dataset

The text dataset comprises over 13,000 naturalistic Chinese sentences constructed using the pest/disease triangle framework. Each sentence integrates three key descriptions: (1) vegetable crops; (2) pest/pathogen features, such as visible symptoms, affected plant parts, and pest size; (3) environmental conditions, including month and humidity. These key

descriptions were combined into coherent and grammatically correct sentences using a template-based generation method, reflecting the descriptive style typically used by farmers.

Pest and Disease Image-Text Identification Model

The architecture of the proposed multimodal (image-text) identification model consists of three main components: an image encoder, a text encoder, and a classifier (Fig. 2). The image encoder is based on YOLOv9 (Wang et al., 2024), which generates multi-scale feature maps that are further processed through a multi-scale feature integration module to obtain a unified visual representation. Simultaneously, the text encoder is based on the RoBERTa-Chinese language model (Liu et al., 2019), fine-tuned to encode diagnostic descriptions that accompany each image. During the training, paired image-text samples are fed into respective encoders. Then, the resulting features are concatenated and passed to a fully connected neural network classifier. Finally, the classifier outputs a probability distribution across 13 predefined pest and disease categories, enabling multi-class predictions with a threshold of 0.5 to determine the final predictions.

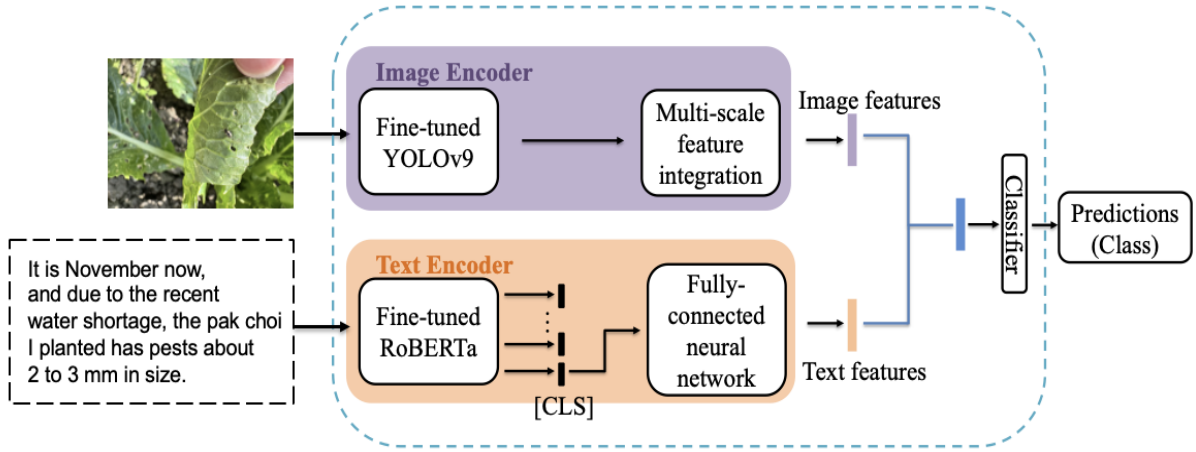

Fig. 2 Pest and disease multimodal identification model structure.

RESULTS & DISCUSSION

The proposed multimodal identification model showed performance improvements when combining image and description inputs, with F1-score increasing from 0.780 (image-only) to 0.870 (image-description) and accuracy improving from 0.795 to 0.918. These results indicate that descriptions of crops and environmental conditions can enhance image-based pest and disease identification. This improvement is illustrated in a representative case where the model with an image input misclassified a pest as a mustard leaf beetle adult, but correctly identified it as a striped flea beetle when description inputs (crop, month, and humidity) were provided alongside the image (Table 1). These outcomes highlight the robustness of multimodal fusion, particularly in addressing classification uncertainties that arise when relying solely on visual information. Despite these improvements, challenges remain for image-only identifications, which exhibited a high quantity of missed detections and misclassifications for three categories (striped flea beetle of 0.305; mustard leaf beetle adult and larva of 0.439 and 0.222, respectively), attributed to insufficient training data and the challenge of differentiating between similar pests. To enable practical application, the multimodal identification model was deployed through a chatbot platform, providing farmers with accessible online identification services.

Table 1 Representative case showing improved identification with description.

Image	Description	Prediction (image)	Prediction (image + description)
	The crop under observation is broccoli, which shows pest infestation. The current observation was made in September under drought conditions.	Mustard leaf beetle adult (Wrong)	Striped flea beetle (Correct)

CONCLUSIONS

This study developed a multimodal identification system for vegetable crop pests and diseases in urban community farming, integrating 2,780 field images with over 13,000 descriptions across 13 categories. The proposed multimodal identification model, combining YOLOv9 for image encoding and RoBERTa for text processing, achieved an F1-score of 0.870 and accuracy of 0.918. Notably, the inclusion of descriptions of crops and environmental conditions enhanced model robustness, particularly in visually ambiguous cases. Future work will refine the dataset and optimize the model structure.

REFERENCES

Wang, C.-Y., Yeh, I.-H., & Liao, H.-Y. M. (2024). YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *Computer Vision – ECCV 2024*, 1-21. https://doi.org/10.1007/978-3-031-72751-1_1

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv*. https://doi.org/10.48550/arXiv.1907.11692