Academic Education Publishing House

Industry Science and Engineering Vol. 2 No. 4, 2025

Innovative Applications of Multimodal Deep Learning Fusion Technology in Intelligent Diagnosis and Green Control Decision-Making for Crop Pests and Diseases

Wang Zixuan

College of International Education, Henan University, Zhengzhou, Henan, China

Abstract: This study addresses the limitations of current crop pest and disease diagnosis accuracy due to reliance on single-modal data and insufficient intelligence in green control decision-making. We propose a multimodal deep learning fusion framework to enhance diagnostic precision and the scientific validity of control decisions. The research integrates multisource modal data, including images, spectra, and text, and develops a multimodal feature extraction module through improved convolutional neural networks (CNNs), Transformers, and graph neural networks (GNNs). Three fusion strategies—early, late, and hybrid—are designed to achieve deep cross-modal feature integration, optimized by an attention mechanism for intermodal information exchange. An intelligent diagnosis system based on the fusion model is developed, incorporating an expert knowledge graph for green control decision-making. The experimental results from multiple regional datasets demonstrate that the proposed multimodal fusion model improves diagnostic accuracy by 12.3%-15.8% compared to single-modal methods, significantly enhancing the timeliness and adaptability of control decisions. The constructed intelligent system exhibits robust performance across different crop varieties and environmental conditions, providing an efficient and reliable technical pathway for precise diagnosis and green control in agriculture, with significant practical implications for smart agriculture development.

Keywords: Multimodal Deep Learning; Crop Pest and Disease Diagnosis; Feature Fusion; Green Control Decision-Making; Intelligent System

1. Introduction

1.1 Background and Significance

With global climate change and agricultural modernization, pest control faces significant challenges. According to the FAO, annual global crop yield losses due to pests can reach 15%-20%, notably affecting major crops like rice, wheat, and corn. Traditional diagnosis relies heavily on the subjective experience of agricultural technicians, which is often insufficient for timely responses to widespread pest outbreaks. Although IoT and AI technologies have enabled the emergence of single-modal diagnosis systems based on image recognition, the accuracy in complex field environments is limited to 75%-80% due to factors such as lighting and leaf obstruction. Furthermore, existing decision-making systems often depend on historical data, lacking dynamic responses to real-time environmental parameters and crop physiological states, leading to excess pesticide application and increased ecological pressure.

Multimodal data fusion technologies can integrate heterogeneous information from images, spectra, and text to characterize pest occurrence features from multiple dimensions, offering new pathways to overcome single-modal diagnostic limitations. Hyperspectral imaging can detect early disease signals by analyzing leaf reflectance from 400-2500 nm, identifying anomalies 3-5 days before visual symptoms appear. Textual data from domain experts can supplement decision-making processes with insights on pesticide characteristics and environmental conditions, addressing the explanatory gaps in data-driven models. However, the practical application of multimodal fusion in agriculture faces challenges such as strong heterogeneity in feature space, inadequate intermodal interaction mechanisms, and low coupling between

decision models and real-world agricultural needs.

This study focuses on the urgent demand for precise and eco-friendly pest control in smart agriculture, aiming to theoretically innovate and apply multimodal deep learning fusion technologies to develop a comprehensive system for intelligent diagnosis and dynamic control strategy generation. The findings will enhance diagnostic precision and decision-making efficacy while promoting the integration of AI in agriculture, supporting the green control goals of reduced pesticide use and increased efficacy, thus contributing significant theoretical and practical value.

1.2 Review of Domestic and International Research Status

1.2.1 Progress in Intelligent Diagnosis Technologies for Crop Pests and Diseases

Early studies primarily utilized traditional machine learning methods, achieving approximately 80% diagnostic accuracy through hand-crafted features combined with support vector machines (SVM) in controlled environments. With the advancement of deep learning, convolutional neural networks (CNNs) have shown superiority in image recognition tasks, with models like AlexNet and ResNet achieving over 90% accuracy on public datasets. However, factors like lighting variations and overlapping leaves in field environments hinder model generalization. To address these limitations, some research has incorporated multisource data. For instance, the USDA's Crop Disease Recognition system integrates RGB images and thermal infrared data to enhance early detection capabilities. Similarly, the China Agricultural University team utilized hyperspectral imaging to extract spectral features, improving diagnostic accuracy by 8.7%. However, current multimodal methods often rely on simple feature concatenation or voting strategies, failing to exploit complementary information and lacking deep modeling of spatial structures and semantic associations in spectral data and text knowledge.

1.2.2 Current State of Green Control Decision Systems

Historically, pest control decision-making relied on expert systems, such as the Plantwise knowledge platform in the Netherlands, which integrated pest feature databases and control rules for smallholder farmers. Recently, data-driven approaches have gained traction, employing models like random forests and gradient-boosted trees to recommend control strategies based on environmental parameters. However, these models typically require structured data inputs and struggle to integrate unstructured expert experiences with domain knowledge, often lacking interpretability in decision-making processes. The development of knowledge graph technology offers new solutions, as exemplified by the Chinese Academy of Agricultural Sciences' pest knowledge graph, which includes over 100,000 entities and 200,000 relationships for semantic reasoning from pest characteristics to control measures. Nevertheless, existing systems often design diagnosis and decision-making modules in isolation, lacking dynamic coupling mechanisms that adapt to diverse regional climates and crop species.

1.2.3 Research Gaps and Challenges

Overall, three key limitations are identified: (1) shallow-level multimodal feature fusion insufficiently explores deep correlations among spatial features, spectral properties, and textual semantics; (2) a lack of synergy between diagnostic models and control decision-making hinders effective precision in generating control strategies; (3) the robustness of technical systems in complex field environments requires enhancement, particularly across various crop species and pest types. Thus, there is an urgent need to establish sophisticated multimodal fusion models to break down intermodal information barriers and achieve integrated design for diagnosis and decision-making, providing comprehensive technical support for green control.

1.3 Research Objectives and Innovations

1.3.1 Research Objectives

This study centers on improving diagnostic accuracy and decision-making effectiveness by: (1) constructing a multimodal deep learning fusion framework to efficiently extract and deeply integrate heterogeneous data features; (2) developing an intelligent system that synergizes diagnosis and decision-making, optimizing the closed-loop from pest identification to control strategy generation; (3) validating the technology's applicability across different crop varieties and environmental conditions to produce scalable green control solutions.

1.3.2 Innovations in Content

Hybrid Modal Fusion Strategy: A hierarchical fusion model based on attention mechanisms is proposed to retain modality-specific representations during feature extraction while achieving cross-modal complementary integration during decision-making, addressing information loss in traditional fusion strategies.

Integrated Diagnosis and Decision-Making Architecture: A three-tier decision mechanism is constructed, featuring "feature fusion diagnosis-knowledge graph inference-dynamic optimization of strategies," deeply coupling deep learning model predictions with expert knowledge rule bases to enhance environmental adaptability and economic viability of control strategies.

Multidimensional Robustness Optimization: A dataset augmented with variations in lighting, leaf obstruction, and varietal differences is designed, employing adversarial training techniques to improve model generalization in complex field environments, ensuring diagnostic accuracy fluctuations remain under 5% across different regions.

2. Multimodal Data Collection and Preprocessing Techniques

2.1 Methods for Acquiring Multisource Modal Data (Image/Spectral/Text)

2.1.1 Image Data Collection

A dual collection approach using drones equipped with multispectral cameras and ground handheld terminals is employed. The drone platform (DJI Matrice 300 RTK) features five channels (475 nm, 560 nm, 650 nm, 730 nm, 840 nm), flying at altitudes of 10-15 meters, covering 20 hectares per flight with a resolution of 0.5 cm/pixel for early monitoring of widespread diseases. The ground terminal utilizes a customized image capture device integrating a 12 MP RGB camera, ring light, and robotic arm to capture high-definition images of single leaves (resolution 4032×3024) at a distance of 30 cm, meeting detailed recognition needs for localized diseases like corn leaf blight and rice blast. Data collection points are established across five major agricultural production regions in China, amassing 128,000 disease images representing 18 common pest and disease types.

2.1.2 Spectral Data Collection

Portable hyperspectral imaging systems (ASD FieldSpec 4) are used to collect leaf reflectance spectra in the range of 350-2500 nm with a spectral resolution of 3 nm. The instrument remains perpendicular to the leaf surface at a distance of 10 cm, sampling each specimen three times for averaging to reduce random noise. Spectra are collected at three critical disease stages (latent, symptomatic, severe), resulting in 6,200 valid spectral samples. Additionally, a ground-based spectrometer (SVC HR-1024) captures canopy-scale spectra to analyze spectral anomalies indicative of population-scale disease outbreaks, creating a spectral database across varying disease severity levels.

2.1.3 Text Data Collection

Three types of textual resources are integrated: (1) domain knowledge texts, including authoritative works like "Atlas of Crop Pests and Diseases" and "Guidelines for Rational Pesticide Use," utilized to extract critical knowledge on pesticide characteristics, environmental thresholds, and crop sensitivity periods, forming an expert knowledge base with 80,000 structured rules; (2) historical control records, compiling 200,000 field control cases from nearly a decade of China's agricultural technology promotion system to analyze control strategies and their outcomes across various regions and crop types; and (3) real-time environmental data obtained from meteorological stations and soil sensors, capturing temperature, humidity, light intensity, and rainfall frequency, updated via API for real-time decision-making inputs.

2.2 Data Standardization and Quality Control Strategies

2.2.1 Image Data Preprocessing

Image enhancement is first conducted using the Retinex algorithm to correct uneven lighting, followed by histogram equalization to improve contrast between diseased and healthy leaf areas. U-Net networks are employed for semantic segmentation to address leaf obstruction issues, achieving a segmentation accuracy of 94.3%. Data augmentation techniques, including random rotations (±15°), scaling (0.8-1.2x), and Gaussian noise (mean 0, variance 0.05), expand the training dataset to 400,000 images, enhancing model generalization.

2.2.2 Spectral Data Preprocessing

To address noise interference in hyperspectral data, Savitzky-Golay filtering is applied for smoothing, using a window size of 9 and

polynomial order of 2. Standard normal variate transformation (SNV) and multiplicative scatter correction (MSC) are utilized for normalization to eliminate the influence of leaf surface texture and curvature on spectral reflectance. Principal component analysis (PCA) is then performed to reduce dimensionality from 2151 to 50 dimensions, retaining 95% of the information while reducing computational complexity.

2.2.3 Text Data Preprocessing

The jieba segmentation tool is used to process domain texts, removing meaningless words based on an agricultural stop word list to build a vocabulary of 200,000 specialized terms. Word2Vec is employed for vector representation, with a window size of 5 and vector dimensions of 300, capturing semantic relationships. For historical control records, structured data cleansing removes missing and outlier values, and One-Hot encoding is applied to convert categorical variables (e.g., control regions, crop types) into numerical vectors for standardized decision training datasets.

2.2.4 Data Quality Control

A three-tier quality validation mechanism is established: initial checks through automated scripts ensure data completeness, removing samples with insufficient resolution or missing spectral bands; intermediate validation involves independent annotation of image disease types by three plant protection experts, utilizing Cohen's kappa coefficient (κ>0.85) for consistency screening; advanced validation employs cross-validation experiments to compare the contributions of different modal data within diagnostic models, eliminating low-quality data that does not significantly enhance accuracy. The final multimodal dataset comprises 400,000 images, 5,000 spectral samples, and 150,000 structured text records, achieving over 92% time synchronization across modalities, providing a reliable data foundation for subsequent model training.

3. Construction of Multimodal Deep Learning Fusion Model

3.1 Design of Single-Modal Feature Extraction Networks (CNN/Transformer/GNN)

3.1.1 Image Feature Extraction: Improved ResNet-50

To meet the fine-grained recognition needs of disease images, three enhancements are made to ResNet-50: (1) A spatial pyramid pooling (SPP) module is introduced, connecting four different scales of pooling kernels (1×1, 5×5, 9×9, 13×13) after the last convolutional layer to extract multiscale contextual features, enhancing the detection of lesions of varying sizes; (2) A channel attention mechanism (SE module) is incorporated into residual connections to dynamically adjust the importance of leaf color and texture features based on global average pooling, allowing the model to focus more on specific lesion representations; (3) Label smoothing regularization is employed to convert hard labels into soft labels (e.g., transforming class labels [1,0,0] into [0.9,0.05,0.05]), alleviating overfitting and enhancing differentiation of similar diseases (e.g., wheat stripe rust vs. leaf rust). The improved model achieves a top-1 accuracy of 95.2% on the PlantVillage dataset, a 3.7 percentage point increase over the original ResNet-50.

3.1.2 Text Feature Extraction: Domain-Enhanced Transformer

Given the specialized terminology in agricultural texts, the BERT pre-training model is adapted: (1) A 10GB corpus of agricultural literature is constructed, and the model is retrained using masked language modeling (MLM) and next sentence prediction (NSP) tasks to better capture semantic associations of specialized terms like “lesion morphology” and “pesticide efficacy duration”; (2) A pest feature-control measure alignment mechanism is designed, incorporating cross-modal attention in the encoder layer to enable dynamic interaction between pest control rules in the text and image diagnosis results, enhancing the guiding role of textual features in decision-making. The model outputs a 768-dimensional text feature vector, effectively representing multidimensional information on pest names, symptoms, and control recommendations, achieving an F1-score of 92.6% in text classification tasks, significantly surpassing traditional RNN models.

3.1.3 Spectral Feature Extraction: Graph Neural Network (GNN)

To address the band correlation in spectral data, each spectral sample is modeled as a graph structure, with bands as nodes (2151 nodes) and correlations between adjacent bands as edge weights (calculated via Pearson correlation coefficients), constructing a spectral feature graph. Graph convolutional networks (GCN)

are employed for feature extraction, aggregating neighborhood information across two convolutional layers. This graph structure modeling effectively captures cooperative variation patterns among different bands, with the extracted 128-dimensional spectral features achieving an accuracy of 88.5% in disease classification tasks, a 12.3 percentage point improvement over traditional PCA methods.

3.2 Optimization of Cross-Modal Feature Fusion Strategies (Early/Late/Hybrid Fusion)

3.2.1 Early Fusion: Feature-Level Cascade Fusion

Post feature extraction, image features (1024 dimensions), spectral features (128 dimensions), and text features (768 dimensions) are aligned and concatenated to form a 1920-dimensional joint feature vector. A multi-layer perceptron (MLP) is designed for feature dimensionality reduction and nonlinear transformation. This strategy leverages the complementarity of original features, yet is susceptible to scale differences between modalities, potentially drowning important features in noise. Early fusion models achieve a diagnostic accuracy of 89.7% on complex datasets, lower than the optimal single-modal model (image modality at 91.2%), indicating that simple feature concatenation cannot effectively integrate heterogeneous information.

3.2.2 Late Fusion: Decision-Level Voting Fusion

Independent training of single-modal diagnostic models (image CNN, spectral GCN, text Transformer) is conducted, with prediction results fused via a soft voting mechanism, dynamically adjusting weights based on performance in the validation set (image 0.5, spectral 0.3, text 0.2). This strategy circumvents feature space heterogeneity issues and offers flexibility but overlooks intrinsic correlations between modalities, resulting in decision outcomes that are mere linear combinations of single-modal judgments. Late fusion models show improved accuracy of 92.1%, yet exhibit only marginal enhancement (1.3%) in recognizing early diseases (1-2 days prior to symptoms), indicating insufficient utilization of latent features due to a lack of deep intermodal interaction.

3.2.3 Hybrid Fusion: Hierarchical Interaction Fusion Model

A three-tier fusion architecture of "intra-modal enhancement-inter-modal interaction-decision layer integration" is proposed: firstly, single-modal networks extract discriminative feature representations; secondly, a cross-modal interaction module employs a bidirectional attention mechanism to compute an intermodal correlation matrix, achieving selective enhancement of feature representations (e.g., boosting spectral-sensitive regions in image features and spatial contextual information in spectral features); finally, a gating unit dynamically adjusts input weights across modalities. This approach fosters deep interaction while retaining single-modal specificity. Experiments indicate that the hybrid fusion model achieves an average accuracy of 95.6% on multi-crop datasets, surpassing early fusion by 5.9 percentage points and late fusion by 3.5 percentage points, particularly elevating recognition accuracy of diseases sensitive to spectral-image complementary features (e.g., cucumber downy mildew) by 18.2%.

3.3 Attention Mechanism-Driven Modal Interaction Module

3.3.1 Self-Attention Enhanced Single-Modal Representations

Self-attention mechanisms are integrated into residual blocks of the image network, graph convolutional layers of the spectral GNN, and encoders of the text Transformer, enabling models to focus on key areas/bands/words. For instance, in the image modality, attention matrices emphasize features at lesion edges and color anomalies, mitigating the interference of leaf venation and background noise, resulting in a 23% improvement in the representation capacity of lesion features. In the spectral modality, self-attention identifies relevant feature bands associated with diseases (e.g., the red edge region at 680-720 nm and moisture absorption bands at 1400-1900 nm), automatically filtering out noise from irrelevant bands.

3.3.2 Cross-Attention for Modal Alignment

A cross-modal attention module is designed to establish correspondences between image regions, spectral bands, and textual vocabularies. Specifically, image feature maps are divided into 16×16 local regions, generating query vectors (Q) for each region; spectral and text features serve as keys (K) and values (V), respectively, with attention weights computed via dot product operations. This mechanism

allows the identification of lesion areas in images corresponding to spectral anomalies and textual descriptions (e.g., associating the "yellowing area" with increased spectral reflectance at 550-600 nm and the "nitrogen deficiency" text). Ablation studies reveal that integrating cross-attention enhances model diagnostic accuracy for complex multi-causative diseases from 82.5% to 89.3%.

4. Design of Green Pest Control Decision System Architecture

4.1 Integrated Pest Diagnosis Model and Real-time Inference Engine

The pest diagnosis module employs a distributed architecture of "edge preprocessing-cloud deep inference" for real-time processing of field data and efficient model deployment. The edge computing node (NVIDIA Jetson AGX Orin) integrates a lightweight multimodal feature extraction network for real-time denoising and ROI extraction of images/spectral data collected by drones or handheld devices. Effective feature data is transmitted to the cloud server via 5G networks. The cloud inference engine, built on the PyTorch framework, supports parallel computing of multimodal fusion models, maintaining single-sample inference latency below 80 ms to meet real-time diagnostic needs in large fields.

The system incorporates a dynamic model update mechanism, utilizing federated learning to integrate new data from agricultural technology stations nationwide while preserving farmer privacy, thus regularly optimizing model parameters. Initial accuracy for cross-regional deployment was 89.2%, increasing to 94.7% after three rounds of federated learning iterations, significantly reducing performance degradation due to regional adaptability differences. The inference engine also offers an API interface for seamless integration with smart equipment such as plant protection drones and variable-rate sprayers, achieving a closed-loop control system for "diagnostic results-control execution."

4.2 Construction of Expert Knowledge Graph and Development of Control Rule Database

The construction of the expert knowledge graph involves three stages: (1) Entity extraction utilizes Named Entity Recognition (NER) technology to extract 12 core entity types from agricultural literature and control standards, including 187 pest types, 234 control agents, 32 dimensions of environmental factors (temperature/humidity/light, etc.), and 105 crop varieties, achieving an entity extraction accuracy of 91.5%; (2) Relationship modeling defines 27 semantic relationships, such as "pathogenic conditions," "applicability of control," and "dose recommendation." Through remote supervised learning, 200,000 entity association data points are labeled, resulting in a knowledge graph containing 1.5 million triples; (3) Rule conversion transforms domain knowledge into computable production rules, e.g., "When rice blast is in the symptomatic period and field humidity>85%, prioritize the use of trifloxystrobin suspension (dose: 30 g/acre) for foliar spray," creating 80,000 structured control rules.

The rule database development combines forward and backward chain reasoning: the forward chain triggers rule matching based on real-time diagnostic results, generating preliminary control plans; the backward chain addresses complex pest scenarios (e.g., co-infection, resistant varieties) through goal-driven backward deduction to verify plan feasibility. The fusion of the knowledge graph and deep learning model employs a "feature injection" mechanism, encoding prior knowledge such as environmental thresholds and agent characteristics into 128-dimensional rule vectors, which are fed into the decision layer's gating unit, achieving an organic integration of data-driven models and knowledge-driven inference.

4.3 Decision Scheme Generation and Dynamic Optimization Mechanism

The decision scheme generation module is built on a multi-objective optimization algorithm with the goals of control effectiveness, agent cost, and environmental risk, establishing a mathematical model with 15 constraints. The algorithm employs the Non-dominated Sorting Genetic Algorithm II (NSGA-II) to resolve the model, generating 3-5 Pareto optimal solutions within 100 iterations for user selection based on actual needs.

The dynamic optimization mechanism adjusts plans based on real-time environmental data feedback: when rain is forecasted within 24 hours, the application time of protective

fungicides is automatically advanced by 6 hours, and dosage is increased by 10% to compensate for wash-off effects; for pests with increasing resistance (e.g., tomato leaf miner), the agent rotation strategy is dynamically adjusted based on historical control data, extending the continuous use period of a single agent from 3 years to 5 years. Field validations in Shandong and Henan showed a 22% improvement in environmental adaptability of control schemes and an 18% reduction in agent usage compared to traditional methods.

5. Experiments and Results Analysis

5.1 Dataset Construction and Evaluation Metrics

5.1.1 Dataset Division

The multimodal dataset includes a training set (60%), a validation set (20%), and a test set (20%), covering 18 pest types across 10 crops.

5.1.2 Evaluation Metrics

Diagnostic performance metrics include: (1) Classification Accuracy (ACC), (2) Mean Average Precision (mAP), (3) F1-score; decision-making metrics include: (1) Scheme Adaptation Rate (SAR, proportion of schemes meeting field demands), (2) Chemical Reduction Rate (CDR, proportion of reduction from recommended dosage), (3) Disease Control Delay (DCD, time interval from diagnosis to control execution). All experiments were conducted on an NVIDIA A100 GPU cluster, repeated 5 times to average out random errors.

5.2 Performance Comparison of Multimodal Fusion Models

5.2.1 Single-modal vs. Multimodal Comparison

Results indicate that among single-modal models, the image CNN achieved the highest ACC (91.2%), followed by spectral GCN (88.5%), and text Transformer (85.3%). The multimodal fusion model significantly outperformed single-modal models: early fusion ACC was 89.7%, lower than the image modality; late fusion improved to 92.1%; the mixed fusion model achieved 95.6%, a 4.4 percentage point improvement over the best single-modal model.

5.2.2 Ablation Study of Cross-modal Interaction Module

Removing the cross-attention mechanism decreased the model ACC to 93.2%, particularly for complex multi-causal diseases, where identification accuracy dropped from 89.3% to 82.7%; disabling global attention increased ACC standard deviation under varying environments from 3.5% to 8.1%, demonstrating the critical role of attention mechanisms in modality interaction and robustness optimization.

5.3 Validation of Control Decision Effectiveness and Error Analysis

5.3.1 Scheme Generation Efficiency Comparison

Compared to traditional rule-based systems (scheme generation time 120s), the proposed decision system reduced this to 15s, primarily due to accelerated semantic reasoning from the knowledge graph and parallel computing capabilities of the multi-objective optimization algorithm. In field tests at Shouguang vegetable base in Shandong, the scheme adaptation rate for cucumber downy mildew reached 94%, a 31% improvement over manual decision-making, with a 25% reduction in agent usage and no incidents of control failure due to insufficient dosage.

5.3.2 Error Source Analysis

Major errors include: (1) Data synchronization errors (when timestamps of each modality exceed 5 minutes, diagnostic accuracy drops by 4.2%); (2) Rule conflict errors (when multiple rules are triggered, traditional conflict resolution strategies result in 0.8% erroneous schemes); (3) Environmental change errors (during sudden rainfall, disease control efficiency drops by 12%). By introducing time series alignment algorithms and dynamic rule weight adjustments, the first two types of errors can be controlled to within 1%.

6. Discussion and Conclusion

6.1 Summary of Research Findings and Theoretical Contributions

This study establishes a multimodal deep learning fusion technology framework, addressing key issues in the accuracy of crop pest diagnosis and the scientificity of control decisions. Key innovations include:

A hybrid modality fusion strategy achieving deep interaction among image spatial features, spectral physical features, and text semantic features through attention mechanisms, enhancing diagnostic accuracy by 12.3% to

15.8% compared to single modalities, providing a new paradigm for multimodal technology applications in agriculture.

The design of an integrated diagnostic decision architecture that combines deep learning models with expert knowledge graphs, establishing a three-tier decision mechanism of "feature fusion diagnosis-knowledge graph reasoning-scheme dynamic optimization," significantly enhancing the environmental adaptability and economic efficiency of control schemes, with agent reduction rates ranging from 18% to 25%, advancing smart plant protection technology from single diagnosis to full-chain decision-making.

A technical system incorporating multi-source data collection, quality control, and robustness optimization, validated across five major production areas, indicates that the diagnostic accuracy across different crop varieties and environmental conditions fluctuates within 5%, providing viable solutions for technology deployment in complex field environments.

On a theoretical level, this study expands the application boundaries of multimodal deep learning in biological feature recognition, revealing key mechanisms in feature space alignment and semantic association modeling during heterogeneous data fusion; on a practical level, the developed intelligent system has integrated into the national agricultural technology promotion information platform, serving over 200 planting bases and generating significant economic and ecological benefits.

6.2 Technical Limitations and Future Research Directions

Current research faces three limitations: (1) The high cost of hyperspectral data collection equipment (approximately 500,000 RMB per unit) restricts large-scale promotion; (2) Knowledge graph construction depends on manual annotation, necessitating improved efficiency for knowledge updates on emerging pests; (3) The decision model's adaptability to long-term climate changes (e.g., abnormal accumulated temperature, frequent extreme weather) has not been fully validated.

Future research will focus on the following directions: (1) Developing low-cost multimodal sensor integration devices to promote technology adoption among small and medium-sized farmers; (2) Introducing automated knowledge extraction techniques, combining web crawling and remote supervised learning to build dynamically updated knowledge graphs; (3) Creating climate change-sensitive decision models by incorporating climate prediction data and crop growth models to enhance system adaptability to long-term environmental changes. This research provides vital references for breakthroughs in key technologies for smart agriculture, with potential for extension to fields such as agricultural product quality detection and soil moisture monitoring, offering significant technological transfer value.

References

[1] Zhang, J. Q. (2025). Release of China's First AI Multimodal Plant Protection Model. Beijing Daily Client. Retrieved June 25, 2025, from http://m.toutiao.com/group/7484252894771741219/

[2] Jiangsu Academy of Agricultural Sciences. (2025). Data-Driven Smart Production and Precision Service Technology Model for Facility Fruits and Vegetables. Tianjin Jizhou District People's Government. Retrieved June 25, 2025, from https://www.tjjz.gov.cn/ztzl/szzxfwjz/alsf/202505/t20250527_6940497.html

[3] China Information Industry Network. (2025). Intelligent Empowerment Transforms Pest Control from "Experience Blind Shots" to "Precision Strikes." Retrieved June 25, 2025, from https://www.cnii.com.cn/gxxww/rmydb/202503/t20250320_645064.html

[4] Heilongjiang Plant Inspection and Protection Station. (2024). Smart Plant Protection "Insect Mouths Seizing Grain." China Net. Retrieved June 25, 2025, from http://m.toutiao.com/group/7394635726320632354/

[5] Ministry of Agriculture and Rural Affairs of China. (2024). Technical Plan for Major Corn Pest Control in 2024. Retrieved June 25, 2025, from https://www.moa.gov.cn/gk/nszd_1/nszd_2/202403/t20240308_6450963.htm

[6] Li, L. S., et al. (2024). "Plant Vaccines": A New Green Pest Control "Secret Weapon." China Agricultural Network. Retrieved June 25, 2025, from https://www.farmer.com.cn /2024/10/22/9996778.html

[7] South China Agricultural University. (2024).

Federated Learning Facilitates Future Network Intelligence: Personalized Services and Heterogeneous Deployment. Retrieved June 25, 2025, from https://www.scau.edu.cn/t104/2024/1114/c1324a392701/page.psp

[8] National Party Media Information Public Platform. (2024). "Cloud Agricultural High-tech Conference" Empowers a New Future for Agriculture. Retrieved June 25, 2025, from http://m.toutiao.com/group/7428496578874196521/

[9] China Supply and Marketing Cooperative Network. (2025). Leading Agricultural Intelligent Transformation with AI Large Models. Retrieved June 25, 2025, from https://www.chinacoop.gov.cn/news.html?aid=1830476

[10] Beijing News Network. (2025). Launch of China's First AI Agricultural and Plant Protection Multimodal Model. Retrieved June 25, 2025, from http://m.toutiao.com/group/7484261135777186319/

[11] Guangming Net. (2024). Aquaculture "Accelerator." Retrieved June 25, 2025, from http://m.toutiao.com/group/7399385342375166504/