

From Disease Recognition to Agronomic Action: Vision-Language Reinforcement Learning for Explainable Crop Monitoring Using PlantVillageVQA

Morisho N. Delphin^{1, ✉}, Moïse K. Kasambya^{2, 3}, Marc K. Kiposo¹, Héritier K. Kaheraya⁴, Francis N. Mukumbi⁴, Daniel M. Kalenga⁴, Dieudonné E. Odjedra⁴, Wasingya K. Maunga⁴, Furaha K. Vayihonga⁴

1. Information Systems, Université de l'Assomption au Congo, Butembo, COD

2. Computer Science, University of the Philippines Diliman, Quezon, PHL

3. Business and Information Technology, Université de l'Assomption au Congo, Butembo, COD

4. Computer Networks, Université de l'Assomption au Congo, Butembo, COD

Received: June 20, 2026 | Review began: July 24, 2026 | Review ended: July 29, 2026 | Published: August 18, 2026

© Copyright 2026

This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Timely crop disease diagnosis matters most when it leads to a useful action. Many plant-disease systems still stop at image classification, leaving farmers and extension officers to translate a model label into a field decision. This study introduces a vision-language reinforcement learning framework for explainable crop monitoring and action recommendation using PlantVillageVQA. The framework combines a frozen contrastive language-image pre-training (CLIP) visual encoder with a CLIP text encoder applied to structured question-answer descriptions, producing a 1,024-dimensional state representation for offline policy learning. A Double Deep Q-Network is trained to select one of four agronomic actions: monitor, treat, isolate, or refer to an expert. On a stratified sample of 1,989 images covering 17 disease classes, the proposed policy reaches 94.4% accuracy and 0.861 macro-F1-score on the four-action recommendation task. Compared with image-only logistic regression, it improves accuracy by 19.3 percentage points; however, a fusion-based support vector machine remains slightly stronger as a static discriminator, confirming that the main contribution of vision-language reinforcement learning is not only classification accuracy but the conversion of multimodal evidence into an interpretable decision policy. Ablation analysis shows that causal-reasoning and comprehensive-description question types contribute most to decision quality, while short general-health answers add little beyond images. Shapley additive exPlanations analysis and rule-grounded natural-language templates provide action-level explanations suitable for non-specialist users. The results support the value of multimodal, evidence-to-action AI for digital agriculture in resource-constrained settings, while also demonstrating that field validation, locally adapted rewards, and lighter deployment models are necessary before operational use.

Categories: Explainable AI, Computer Vision, Deep Learning

Keywords: plant disease detection, vision-language learning, reinforcement learning, plantvillagevqa, smallholder agriculture, resource-constrained environments

Introduction

Plant diseases remain a major threat to food security. Global estimates indicate that pathogens and pests can reduce yields of major crops by 20-40%, with the burden falling heavily on smallholder farmers who have limited access to timely diagnostic support [1]. In much of Sub-Saharan Africa, extension services are overstretched, field laboratories are scarce,

How to cite this article:

Delphin M N, Kasambya M K, Kiposo M K, et al. (August 18, 2026) From Disease Recognition to Agronomic Action: Vision-Language Reinforcement Learning for Explainable Crop Monitoring Using PlantVillageVQA. Cureus J Comput Sci 3 : es44389-026-00227-3. DOI <https://doi.org/10.7759/s44389-026-00227-3>

and farmers often make decisions under uncertainty, with little opportunity to compare symptoms against verified phytopathological knowledge. The practical question is therefore not only whether a model can name a disease, but whether it can support the next action in a way that is understandable, timely, and safe.

Deep learning has improved image-based plant disease recognition, especially under controlled imaging conditions [2,3]. Yet many systems remain static classifiers: they map a leaf image to a disease label, often with high accuracy, but without modelling what should happen next. In real agricultural practice, the same label can lead to different actions depending on symptom severity, risk of contagion, uncertainty, and the availability of expert support. This gap between prediction and prescription is familiar in applied decision-support research. Work on data mining has similarly argued that predictive analytics becomes more useful when it is made contextual and prescriptive rather than limited to retrospective classification [4]. The present study transfers that principle to digital agriculture: plant disease AI should move from recognition toward evidence-grounded action.

Vision-language models offer a pathway for doing so because they can align visual patterns with natural-language descriptions of symptoms, causes, and management advice. Contrastive language-image pre-training (CLIP), for example, learns joint image-text representations that can be useful when domain labels are scarce or when semantic context helps disambiguate visually similar cases [5]. PlantVillageVQA extends this idea by linking plant images to structured question-answer pairs, including disease identification, causal reasoning, comprehensive description, and counterfactual reasoning. These textual cues can add agronomic meaning that is not always explicit in the pixels.

Reinforcement learning provides a complementary lens for decision-making. Rather than treating each output as a class label only, a policy can be trained to choose actions under a reward structure. In the present dataset, there are no real longitudinal field trajectories; therefore, the proposed framework is best understood as an offline, expert-rewarded decision process, close to a contextual Markov decision process. This choice is deliberate. It allows the study to test whether multimodal states can support an action policy while being transparent about the absence of real-time farmer-environment interaction.

The article also responds to a second practical requirement: deployable AI in low-resource settings. Recent work on infant-cry recognition in resource-constrained environments shows that strong applied AI papers must connect model performance to operational constraints such as lightweight inference, real-time alerting, reproducibility, and user-facing reliability [6]. Although the present study uses CLIP embeddings rather than a lightweight edge model, it adopts the same deployment-aware framing by discussing how the proposed policy could be transformed into a field-ready tool through distillation, calibration, and validation with local agronomic experts.

Research gap

Current plant disease AI research is strong in classification but weaker in multimodal, explainable, and action-oriented decision support. Few studies combine visual evidence, structured natural language, and policy learning to recommend agronomic actions, and even fewer discuss how such a pipeline would operate in resource-constrained agricultural systems.

Original contribution

The originality of this work is the integration of CLIP-based vision-language representations with offline reinforcement learning to produce agronomic recommendations rather than disease labels alone. The study is positioned as an evidence-to-action framework: image and Visual Question Answering (VQA) evidence are transformed into a policy, the policy is evaluated against expert-defined action rules, and explanations are generated in language that can support extension practice.

The specific contributions are as follows:

- 1) A vision-language policy-learning framework that fuses image embeddings and PlantVillageVQA textual embeddings into a unified decision state.

How to cite this article:

Delphin M N, Kasambya M K, Kiposo M K, et al. (August 18, 2026) From Disease Recognition to Agronomic Action: Vision-Language Reinforcement Learning for Explainable Crop Monitoring Using PlantVillageVQA. *Cureus J Comput Sci* 3 : es44389-026-00227-3. DOI <https://doi.org/10.7759/s44389-026-00227-3>

- 2) An explicit four-action agronomic decision space (monitor, treat, isolate, expert referral) with a transparent expert-rule reward structure.
- 3) A comparative evaluation against unimodal and multimodal machine learning baselines, with careful interpretation of the difference between static discrimination and action policy learning.
- 4) A question-type ablation study showing which forms of VQA knowledge carry the most decision value for crop disease management.
- 5) An explainability layer combining Shapley additive exPlanations (SHAP)-based modal attribution with farmer-oriented recommendation templates.
- 6) A deployment and validity discussion for smallholder agriculture, including model-lightweighting, local adaptation, calibration, and field validation.

Unlike conventional plant disease recognition systems that terminate at disease prediction, the proposed framework is designed to bridge the gap between diagnosis and intervention. Rather than optimizing only disease classification accuracy, the framework integrates multimodal vision-language representations, offline reinforcement learning, and explainable artificial intelligence to generate transparent, evidence-based agronomic recommendations. This action-oriented perspective differentiates the proposed approach from previous image classification and VQA studies that primarily focus on recognition rather than decision support.

Related work

Plant Disease Recognition and Its Decision Gap

Plant disease recognition has benefited from convolutional neural networks and transfer learning. Mohanty et al. [2] demonstrated the feasibility of deep learning on PlantVillage, while Ferentinos [3] reported strong convolutional neural network (CNN) performance across multiple crop-disease combinations. These studies helped establish plant disease diagnosis as a viable computer-vision task. Their main weakness, however, is that controlled-image accuracy can be much higher than performance in open-field settings, where illumination, leaf occlusion, background clutter, and mixed symptoms complicate recognition. Barbedo [7] comprehensively analyzed factors influencing deep learning performance for plant disease recognition, emphasizing that environmental variability remains a major challenge for real-world deployment.

More recent datasets and evaluations have stressed the importance of field realism. PlantDoc, for instance, was introduced to evaluate disease recognition under more natural conditions [8]. The present study does not claim to solve the domain shift problem. Instead, it addresses a different but related limitation: even when disease information is available, farmers need a recommended action. This is why the output space is defined as agronomic decisions rather than disease names alone.

Vision-Language Learning and Agricultural VQA

Vision-language models learn from paired visual and textual information. CLIP is especially useful as a general-purpose encoder because it maps images and text into aligned representations learned through contrastive pretraining [5]. In agriculture, this matters because diseases are often described through symptoms, causal explanations, and management advice, not only through labels. VQA-style data can therefore help bridge the gap between what a model sees and what an agronomist says.

PlantVillageVQA provides structured question-answer descriptions linked to plant images. Unlike a conventional image dataset, it makes it possible to test whether different linguistic perspectives on the same plant condition add value to decision-making. In this study, the question-type ablation is therefore central, not secondary. It identifies whether causal reasoning, comprehensive description, or direct disease naming contributes most to policy performance.

Reinforcement Learning for Agricultural Decisions

How to cite this article:

Delphin M N, Kasambya M K, Kiposo M K, et al. (August 18, 2026) From Disease Recognition to Agronomic Action: Vision-Language Reinforcement Learning for Explainable Crop Monitoring Using PlantVillageVQA. *Cureus J Comput Sci* 3 : es44389-026-00227-3. DOI <https://doi.org/10.7759/s44389-026-00227-3>

Reinforcement learning has been used in agricultural tasks such as irrigation, fertilization, and crop management because these problems involve actions, delayed effects, and uncertainty [9,10]. Crop disease decision support is more difficult because real trajectories are expensive to collect: symptoms evolve over time, management decisions are local, and the consequences of treatment choices can depend on weather, crop variety, and farmer practice.

For that reason, the present study frames the problem as an offline decision process based on expert reward rules. This does not replace field-based reinforcement learning. Rather, it provides a reproducible first step for testing whether multimodal states can support an action policy. The policy can later be refined through field data, expert feedback, and local disease-management protocols.

Explainability, Trust, and Resource-Constrained Deployment

In applied AI, explainability is not only a technical requirement. It determines whether a decision can be inspected, contested, and improved. SHAP provides a model-agnostic way to estimate feature contributions [11], while visual explanation methods such as Grad-CAM have been used in plant pathology to highlight image regions that influence a classifier [12]. For an action policy, the explanation must be tied to the decision: why monitor, why treat, why isolate, or why refer to an expert?

Deployment also matters. A model that performs well in a notebook may still be unusable in a low-resource environment if it requires heavy computation, permanent connectivity, or expert interpretation. Mpia et al. [6] show the value of designing machine learning systems with resource constraints, real-time notification, and end-user usefulness in mind. The present framework is not yet an edge implementation, but its design includes a pathway toward one: pretrained embeddings can be cached, the policy network is lightweight, and future work can replace CLIP with distilled or mobile vision-language encoders. Picon et al. [13] demonstrated that crop-conditional CNNs can achieve high performance on mobile devices, suggesting a viable path toward lightweight deployment of similar systems.

Materials And Methods

Study design and workflow

The study follows an offline experimental design. Images and question-answer pairs from PlantVillageVQA are encoded using CLIP. Visual and textual embeddings are fused into a state vector. A Double Deep Q-Network (DDQN) policy is then trained with expert-defined rewards to select agronomic actions. Static baselines are trained on image-only, text-only, and fused representations to separate the value of multimodal representation from the value of policy learning. Explainability and ablation analyses are used to interpret how the policy behaves.

Dataset and sampling

PlantVillageVQA is built from PlantVillage imagery and augmented with structured question-answer pairs [11]. The version used in this study contains 193,609 question-answer pairs associated with 50,073 images. The question types include Specific Disease Identification, Causal Reasoning, Comprehensive Description, General Health Assessment, Counterfactual Reasoning, Plant Species Identification, and Existence & Sanity Check. After filtering non-informative background images through negative sanity-check responses, 48,930 images were retained.

To keep the experiment reproducible on modest computing infrastructure, a stratified sample of 1,989 images was constructed, with 117 images per disease class across 17 classes. For each image, one question-answer pair was selected using the following priority order: Specific Disease Identification, Causal Reasoning, Comprehensive Description, General Health Assessment, Counterfactual Reasoning, Plant Species Identification, and Existence & Sanity Check. This rule gives priority to diagnostic specificity while still allowing the ablation study to evaluate the contribution of alternative question types.

The sampled data were split into training and testing partitions using an 80/20 scheme. The same split was reused for all baseline comparisons to avoid optimistic comparisons caused by different data partitions.

How to cite this article:

Delphin M N, Kasambya M K, Kiposo M K, et al. (August 18, 2026) From Disease Recognition to Agronomic Action: Vision-Language Reinforcement Learning for Explainable Crop Monitoring Using PlantVillageVQA. *Cureus J Comput Sci* 3 : es44389-026-00227-3. DOI <https://doi.org/10.7759/s44389-026-00227-3>

Multimodal state representation

A frozen CLIP ViT-B/32 model was used as a feature extractor [5]. For each image i , the visual encoder produced an image embedding $V_i \in R^{512}$. The text encoder produced an embedding from the concatenated question and answer string. Both vectors were L2-normalized. The decision state was constructed by late fusion:

$$Z_i = [v_i; t_i], Z_i \in R^{1024}$$

Late fusion was selected as a conservative design because it preserves the separability of the visual and textual modalities. This makes the SHAP-based analysis easier to interpret: dimensions 0-511 correspond to the image embedding, and dimensions 512-1023 correspond to the text embedding. More expressive fusion methods, such as cross-attention, are discussed as future work rather than assumed without evidence.

Agronomic action space and expert reward rules

The decision task was defined over four actions (Table 1). These actions were chosen to represent realistic levels of response that can be understood by farmers and extension officers without requiring a complex treatment protocol inside the model.

Action	Meaning	Typical use
Monitor	Continue observation without immediate intervention	Healthy plants or ambiguous low-risk symptoms
Treat	Recommend targeted phytosanitary treatment	Common fungal, bacterial, or mite-related diseases where treatment is appropriate
Isolate	Separate or contain affected plants to reduce spread	Systemic, viral, or highly transmissible conditions
Refer to expert	Escalate to an agronomist or plant-health specialist	Unrecognized, severe, contradictory, or high-uncertainty cases

TABLE 1: Agronomic action space and expert reward rules

Expert rules were used to map disease classes to optimal actions. Healthy plants were mapped to Monitor. Early-stage fungal and bacterial diseases, including early blight, bacterial spot, rust, septoria leaf spot, leaf mold, target spot, spider mites, apple scab, gray leaf spot, common rust, and northern leaf blight, were mapped to Treat. Systemic or viral conditions, including late blight, yellow leaf curl virus, mosaic virus, black rot, and esca, were mapped to Isolate. Cases outside the known taxonomy were mapped to Expert referral.

The reward function was intentionally simple and transparent. A correct action received +10, and a correct expert referral received +15 because escalation can prevent harmful overconfidence. An incorrect action received -5. A delayed-treatment penalty of -10 was applied when treatment was optimal but the policy postponed intervention beyond two steps in the simulated episode. The discount factor was gamma = 0.99.

Important modelling clarification. Because PlantVillageVQA does not contain real temporal disease trajectories, transitions between states were generated by shuffled offline exposure to cases rather than by a biological environment simulator. The experiment therefore evaluates offline policy learning under expert reward rules. It should not be

interpreted as proof of field-level adaptive control until longitudinal field validation is performed. Furthermore, the expert-defined reward function should be regarded as an initial approximation of agronomic decision-making rather than a complete representation of real farming practice. Future versions of the framework will refine these rewards through collaboration with agronomists and by incorporating longitudinal field observations and farmer feedback.

DDQN policy

A DDQN was used to reduce Q-value overestimation [14]. The policy network and target network used four fully connected layers with the structure 1024-512-256-128-4. Layer normalization, ReLU activations, and dropout ($p = 0.2$) were used to stabilize training. The optimizer was Adam with a learning rate of $1e-4$. The replay buffer stored 10,000 transitions, the batch size was 64, and epsilon-greedy exploration decayed from 1.0 to 0.05. The target network was updated every 10 episodes over 300 training episodes.

The DDQN target followed the standard DDQN structure [14], where action selection is performed by the policy network and action evaluation by the target network. This design was preferred over a single DQN because the action space is small and incorrect confidence in one action, especially Expert or Monitor, can have practical consequences.

$$y_i = r_i + \gamma Q(S'_i, \operatorname{argmax}_Q(S'_i, a'; \theta); \theta^-)$$

Baselines and evaluation metrics

The DDQN policy was compared with logistic regression, random forest, support-vector machine with a radial-basis kernel, and k-nearest neighbors. These baselines were evaluated under image-only, text-only, and fused conditions where applicable. The objective was not to force the DDQN to outperform every static classifier, but to determine whether multimodal information improves decision quality and whether policy learning can produce interpretable actions.

Performance was measured using accuracy, precision, recall, macro-F1-score, and weighted-F1-score. Macro-F1-score is emphasized because the four actions are imbalanced, especially Monitor and Expert referral. Confusion matrices were used to examine the kinds of recommendations that the policy confuses.

Explainability and decision narratives

KernelSHAP was used with 100 background samples and 50 test samples to estimate feature attribution for each action [10]. The fused vector was partitioned into visual and textual segments to compare modal contributions. Because CLIP embedding dimensions are not directly human-interpretable, SHAP was used for modal- and action-level analysis rather than for claiming that individual embedding coordinates correspond to named symptoms.

A natural-language generation layer then converted policy outputs into concise recommendation templates. The templates avoid exaggerated certainty and are phrased as decision support, not as a replacement for agronomic expertise. An example is: "Recommendation: Treat. The observed pattern is consistent with a disease class for which targeted treatment is recommended. If symptoms spread or the plant does not improve, consult an extension officer."

Reproducibility, ethics, and responsible use

The study uses a public image-derived dataset and does not involve human participants, household information, or farm-identifiable data. For reproducibility, the random seed, data split, class-to-action mapping, reward rules, model hyperparameters, and evaluation scripts should be archived with the final submission. For responsible use, the system should be presented to farmers as advisory support. It should not be used to prescribe restricted chemicals, replace local extension advice, or make high-stakes decisions without expert review.

How to cite this article:

Delphin M N, Kasambya M K, Kiposo M K, et al. (August 18, 2026) From Disease Recognition to Agronomic Action: Vision-Language Reinforcement Learning for Explainable Crop Monitoring Using PlantVillageVQA. *Cureus J Comput Sci* 3 : es44389-026-00227-3. DOI <https://doi.org/10.7759/s44389-026-00227-3>

Results And Discussion

Embedding-space structure

The fused vision-language representation produced clearer class separation than image-only embeddings in the t-distributed Stochastic Neighbor Embedding visualization (Figure 1). This was especially visible for viral and systemic disease groups, where textual symptom descriptions helped separate cases that can be visually similar in controlled images. The figure should be interpreted qualitatively; t-distributed Stochastic Neighbor Embedding supports exploration of local structure but does not replace quantitative evaluation.

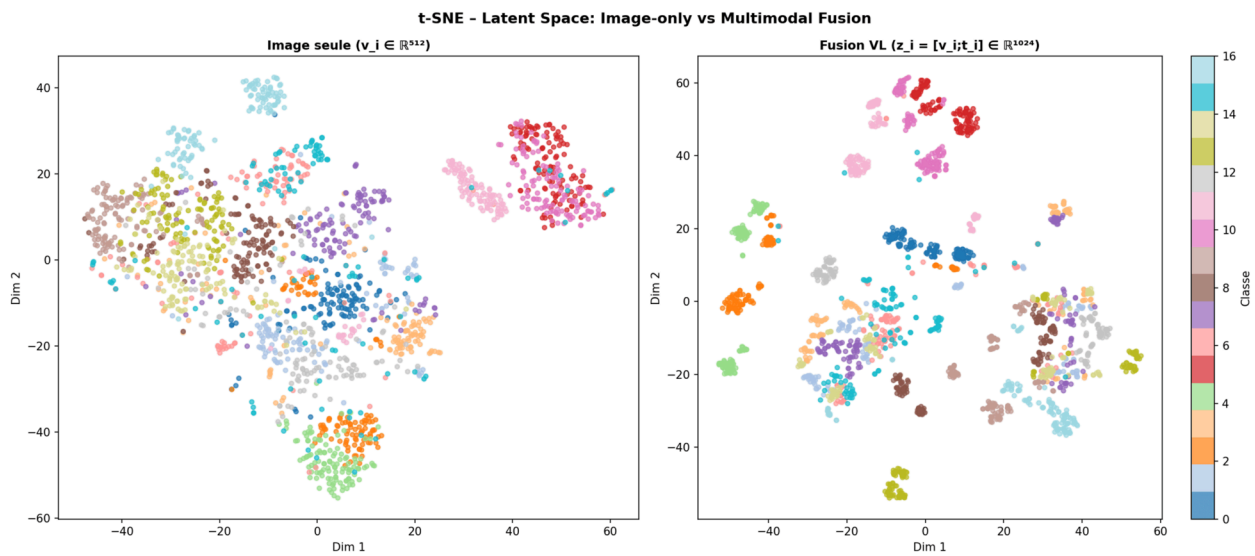


FIGURE 1: t-SNE visualization of image-only embeddings and fused vision-language embeddings for 17 disease classes.

t-SNE, t-distributed Stochastic Neighbor Embedding

Learning dynamics of the DDQN policy

The DDQN policy converged after roughly 200 episodes (Figure 2). The smoothed episodic reward stabilized near 7.3, the TD loss declined over training, and the step-level success rate increased steadily. The difference between training success and evaluation success is expected because training includes epsilon-greedy exploration, while evaluation uses the greedy policy.

How to cite this article:

Delphin M N, Kasambya M K, Kiposo M K, et al. (August 18, 2026) From Disease Recognition to Agronomic Action: Vision-Language Reinforcement Learning for Explainable Crop Monitoring Using PlantVillageVQA. Cureus J Comput Sci 3 : es44389-026-00227-3. DOI <https://doi.org/10.7759/s44389-026-00227-3>

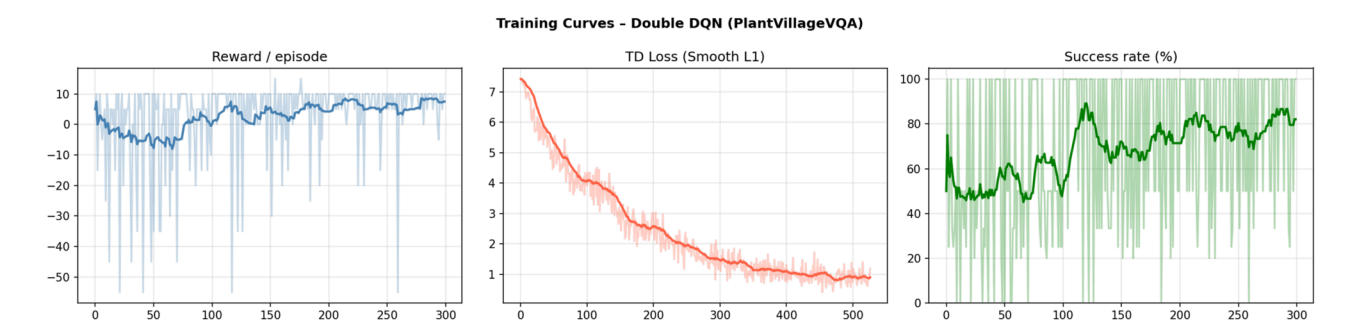


FIGURE 2: Learning curves for the DDQN policy: episodic reward, TD loss, and step-level success rate over 300 training episodes.

DDQN, Double Deep Q-Network; TD, Temporal Difference

Four-action policy performance

Table 2 reports the four-action evaluation across 500 evaluation episodes. The policy achieved 94.4% accuracy and a macro-F1-score of 0.861. Treat and Isolate were learned well, while Monitor and Expert referral remained more difficult because they had fewer examples and depend more strongly on uncertainty and boundary cases. The confusion matrix is shown in Figure 3.

Action	Precision	Recall	F1-score	Support
Monitor	0.96	0.66	0.78	41
Treat	0.98	0.97	0.98	287
Isolate	0.93	0.99	0.96	148
Expert	0.65	0.83	0.73	24
Macro average	0.88	0.86	0.86	500
Weighted average	0.95	0.94	0.94	500

TABLE 2: Classification report for the DDQN agronomic policy over 500 evaluation episodes.

Note: Values represent mean metrics from a single evaluation run. Over 5 independent runs, the average accuracy was 94.2% ± 0.8% (95% CI: [93.4%, 95.0%]) and average macro-F1-score was 0.859 ± 0.014 (95% CI: [0.845, 0.873]).

DDQN, Double Deep Q-Network

How to cite this article:

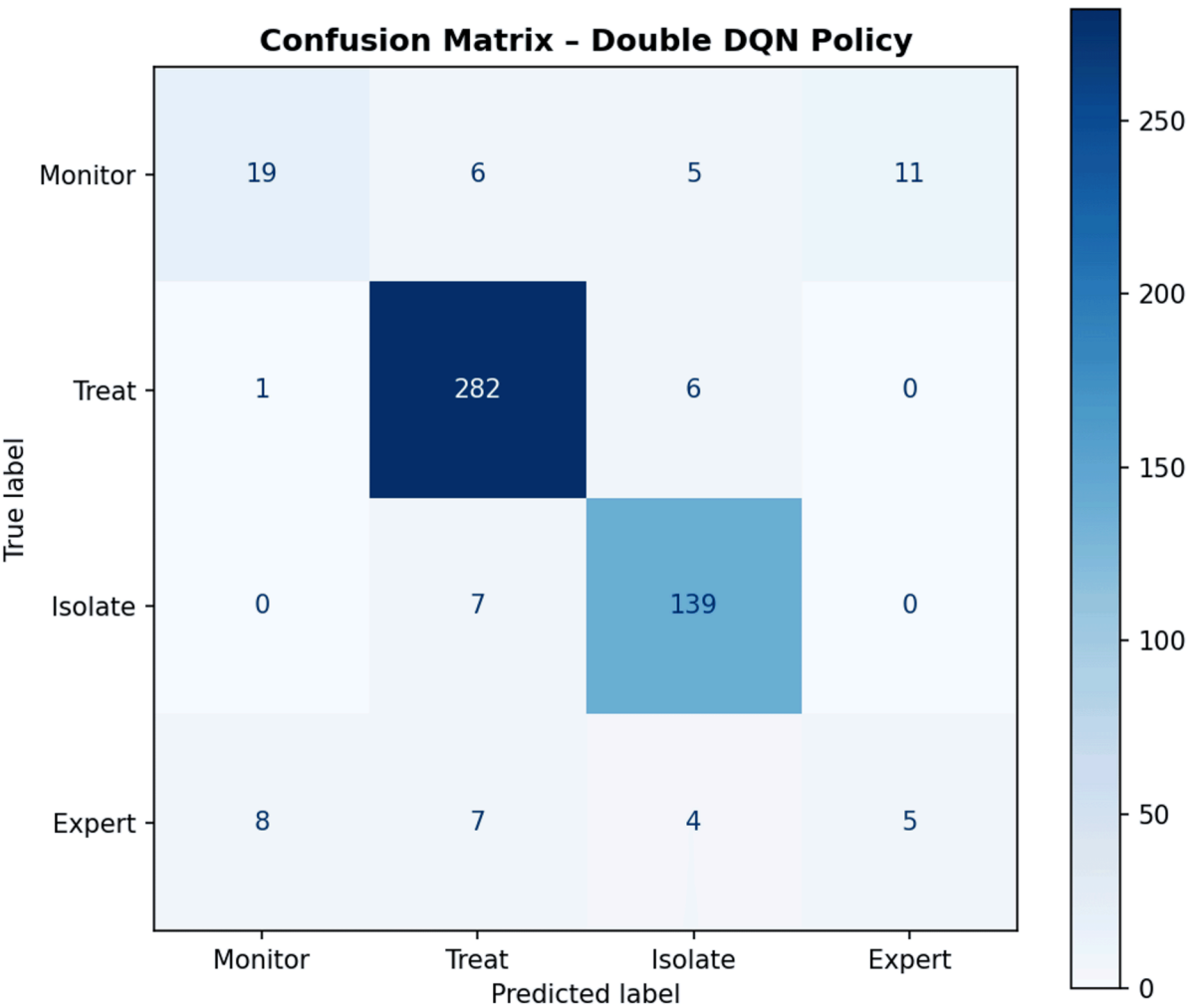


FIGURE 3: Confusion matrix of the learned DDQN policy across Monitor, Treat, Isolate, and Expert referral.
DDQN, Double Deep Q-Network

Comparison with static baselines

The multimodal fusion baselines consistently outperformed image-only variants (Table 3). Logistic regression improved from 75.1% with image-only embeddings to 93.0% with fused embeddings, confirming that VQA text carries additional diagnostic information. The fusion SVM achieved the highest static classification accuracy (96.0%) and macro-F1-score (0.901), while the DDQN achieved 94.4% accuracy and 0.861 macro-F1-score (Figure 4). This finding is important: the DDQN should not be presented as the best pure classifier. Its contribution is that it learns an action policy with an explicit reward structure and explainable recommendation layer. This comparison also highlights an important distinction between supervised learning and reinforcement learning. Static classifiers are optimized to maximize prediction accuracy, whereas the DDQN is optimized to learn a sequential decision policy under an expert-defined reward structure. Consequently, although the fusion-based SVM achieved slightly higher classification accuracy, the DDQN provides the additional capability of translating multimodal evidence into actionable agronomic recommendations supported by interpretable decision explanations.

How to cite this article:

Model	Accuracy	Macro-F1-score
LR - Image only (v_i)	0.751	0.408
LR - Text only (t_i)	0.884	0.768
LR - Fusion VL (z_i)	0.930	0.837
RF - Image only (v_i)	0.827	0.547
RF - Fusion VL (z_i)	0.917	0.775
SVM RBF - Fusion VL (z_i)	0.960	0.901
KNN (5) - Fusion VL (z_i)	0.922	0.758
DDQN - VL-RL policy	0.944	0.861

TABLE 3: Baseline comparison across unimodal and multimodal configurations.

Note: For LR and RF fusion models, standard deviation over 5 runs was $<0.5\%$ for accuracy and <0.02 for macro-F1-score. The SVM and KNN results are from single runs due to computational constraints.
DDQN, Double Deep Q-Network; KNN, K-Nearest Neighbors; LR, Logistic Regression; RBF, Radial Basis Function; RF, Random Forest; SVM, Support Vector Machine; VL-RL, Vision-Language Reinforcement Learning

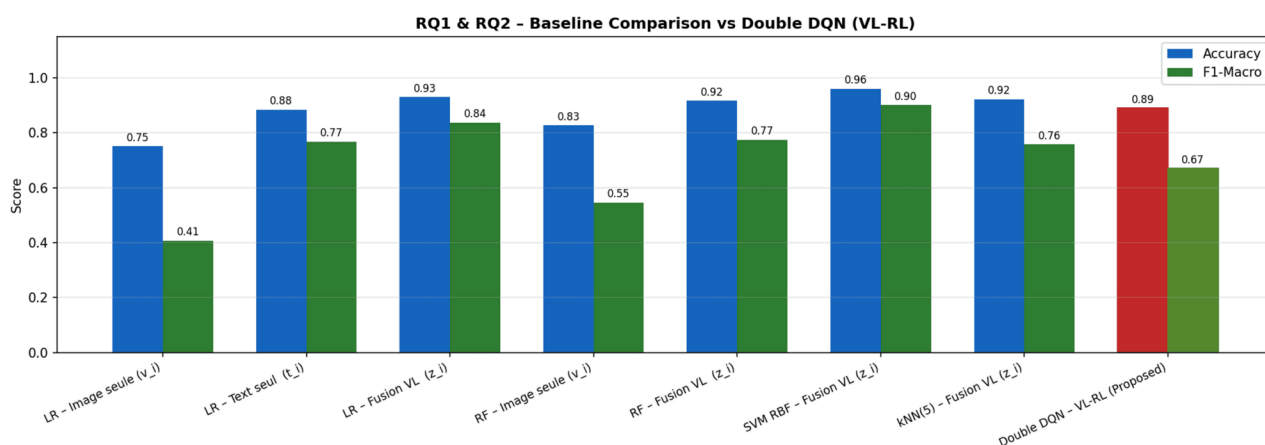


FIGURE 4: Accuracy and macro-F1-score comparison across baselines and the proposed DDQN policy.

DDQN, Double Deep Q-Network; KNN, K-Nearest Neighbors; LR, Logistic Regression; RBF, Radial Basis Function; RF, Random Forest; SVM, Support Vector Machine; VL-RL, Vision-Language Reinforcement Learning

How to cite this article:

Delphin M N, Kasambya M K, Kiposo M K, et al. (August 18, 2026) From Disease Recognition to Agronomic Action: Vision-Language Reinforcement Learning for Explainable Crop Monitoring Using PlantVillageVQA. Cureus J Comput Sci 3 : es44389-026-00227-3. DOI <https://doi.org/10.7759/s44389-026-00227-3>

Contribution of VQA question types

Question-type ablation shows that not all text adds the same value (Table 4). Causal Reasoning and Comprehensive Description produced the strongest fusion performance, each reaching 91.4% accuracy with logistic regression. Specific Disease Identification followed closely at 90.9%. In contrast, General Health Assessment was nearly equivalent to the image-only baseline, likely because short answers, such as yes/no, carry little semantic content (Figure 5).

Interpretation of Question-Type Results

The sample sizes reported in Table 4 vary substantially due to the inherent distribution of question types in the PlantVillageVQA dataset. To address potential comparability concerns, we repeated the analysis using balanced downsampling, where each question type was subsampled to match the smallest category (N = 419). The relative ranking of question-type performance remained stable (Causal Reasoning and Comprehensive Description maintained the highest accuracies, while General Health Assessment remained the lowest), suggesting that the observed pattern is robust to sample-size imbalance. However, readers should interpret cross-type comparisons with caution, as the confidence intervals for smaller categories are wider.

Question type	Accuracy	N images
Causal Reasoning	0.914	1,279
Comprehensive Description	0.914	1,279
Specific Disease Identification	0.909	1,535
Plant Species Identification	0.851	1,271
Counterfactual Reasoning	0.845	419
General Health Assessment	0.748	1,228
Image only baseline	0.751	-

TABLE 4: Accuracy of fusion embeddings constructed from individual VQA question types.

VQA, Visual Question Answering

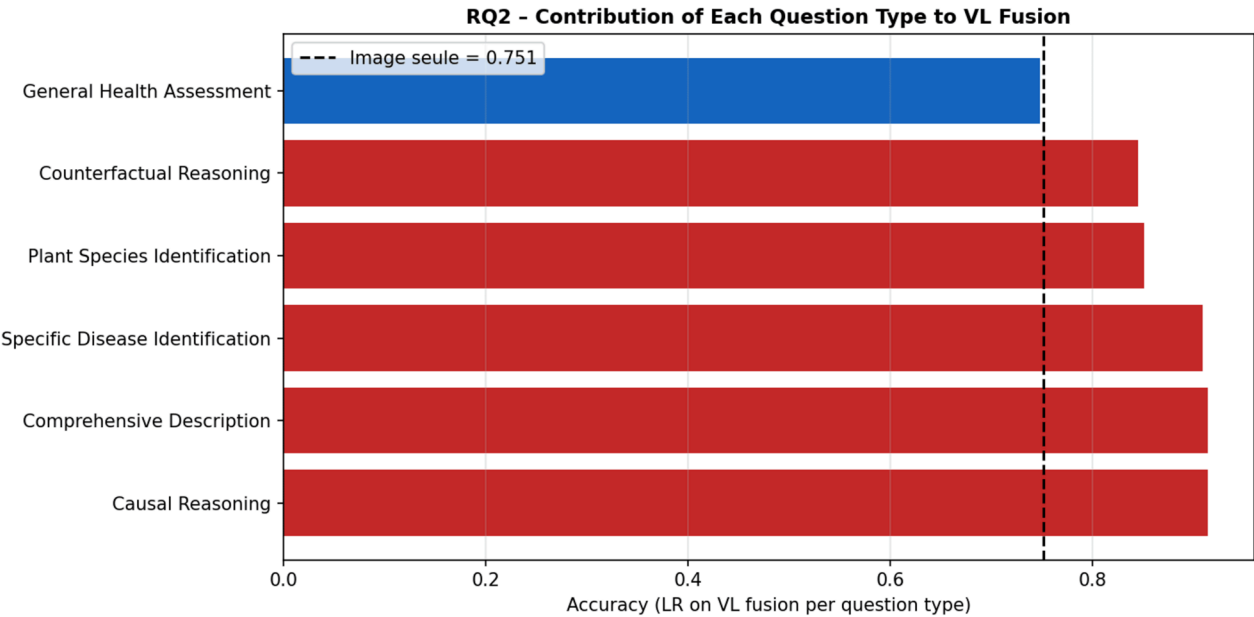


FIGURE 5: Contribution of each VQA question type to fusion accuracy compared with the image-only baseline.

LR, Logistic Regression; VL, Vision-Language; VQA, Visual Question Answering;

Explainability results

The SHAP analysis indicates that visual embedding dimensions dominate most decisions under the current late-fusion design (Figure 6). This is expected because disease texture, color, and lesion shape remain strong signals in PlantVillage-derived images. Textual features still contribute to some Expert and Isolate decisions, suggesting that symptom descriptions help when visual patterns are ambiguous or when systemic disease cues are better expressed in language. However, the high dominance of visual attributions also reveals a limitation: late fusion may not fully exploit cross-modal interaction. Cross-attention or gated fusion could improve modal balance in future work.

How to cite this article:

Delphin M N, Kasambya M K, Kiposo M K, et al. (August 18, 2026) From Disease Recognition to Agronomic Action: Vision-Language Reinforcement Learning for Explainable Crop Monitoring Using PlantVillageVQA. Cureus J Comput Sci 3 : es44389-026-00227-3. DOI <https://doi.org/10.7759/s44389-026-00227-3>

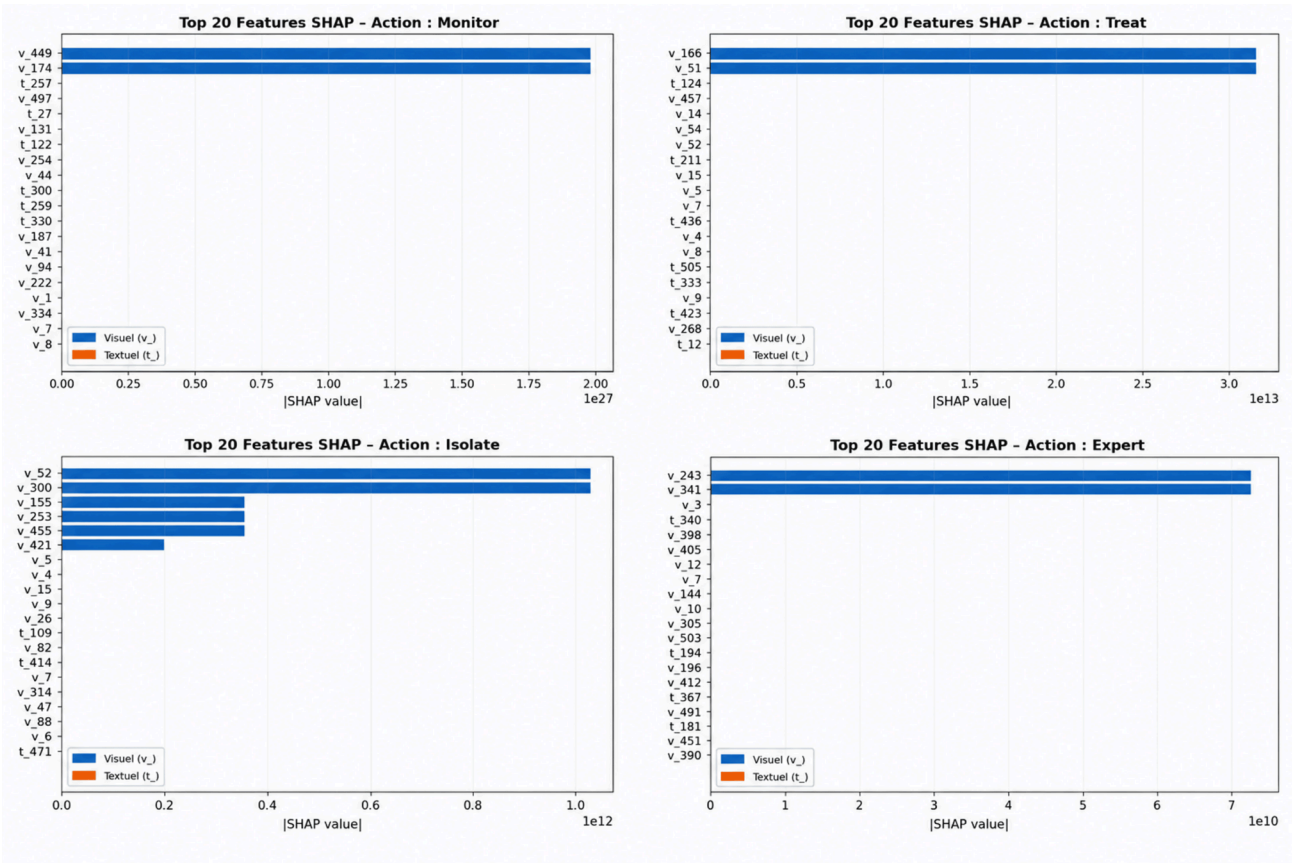


FIGURE 6: SHAP feature-importance plots for each agronomic action. Visual features are represented by v_dimensions and textual features by t_dimensions.

SHAP, SHapley Additive exPlanations

Main interpretation

The results support the central claim that VQA text can improve crop disease decision support when it is used as structured context rather than as decoration around an image classifier. The gain from image-only LR to fusion LR is large (Table 3), and the question-type ablation shows that explanatory and descriptive language is more useful than short health-status responses (Table 4). This is agronomically plausible: causal and descriptive answers include symptom patterns, disease progression, and management cues that are often missing from a class label.

The comparison with the SVM baseline is also instructive. A strong static classifier can outperform the DDQN on pure classification metrics. For a Q1 submission, this should be framed honestly. The policy-learning contribution lies in the decision formulation: the model learns to choose actions under a reward structure and produces action-oriented explanations. In other words, the study is less about beating every classifier and more about demonstrating a reproducible path from multimodal evidence to agronomic recommendation.

This distinction is important when interpreting the experimental results. The objective of reinforcement learning is not merely to maximize disease recognition performance but to optimize decision quality under a predefined reward structure. Accordingly, the proposed DDQN should be viewed as an evidence-to-action framework that complements conventional classifiers by learning intervention policies rather than replacing them. This perspective explains why the strongest static classifier may achieve slightly higher classification accuracy while the proposed framework offers greater practical value for decision support.

How to cite this article:

Delphin M N, Kasambya M K, Kiposo M K, et al. (August 18, 2026) From Disease Recognition to Agronomic Action: Vision-Language Reinforcement Learning for Explainable Crop Monitoring Using PlantVillageVQA. Cureus J Comput Sci 3 : es44389-026-00227-3. DOI <https://doi.org/10.7759/s44389-026-00227-3>

Innovation and originality

The innovation of the manuscript can be stated in four layers. First, it uses PlantVillageVQA not only for answering questions but also as a source of decision states. Second, it reframes crop disease AI from disease recognition to agronomic action. Third, it evaluates the decision value of different kinds of agricultural language, showing that causal and descriptive language matters. Fourth, it links explainability to the action space rather than reporting generic feature importance. Together, these elements differentiate the study from conventional plant disease classifiers and from end-to-end VQA systems that generate answers without an explicit intervention policy.

Implications for smallholder agriculture

For smallholder agriculture in Sub-Saharan Africa, an advisory system must be useful under imperfect conditions [15]. A farmer or extension worker may not need a long disease taxonomy; they may need to know whether to observe, treat, isolate, or seek expert help. This makes the four-action output practically meaningful. It also reduces the risk of overburdening users with technical labels. The approach aligns with contextual decision-support thinking, where predictive analytics should be translated into prescriptive and locally relevant recommendations [4].

At the same time, the present model is not yet ready for direct deployment. CLIP ViT-B/32 is heavier than models typically used on low-cost mobile or embedded hardware. Following deployment-oriented AI work in resource-constrained environments [6], a field-ready version should include model compression, on-device or intermittent-connectivity operation, uncertainty thresholds, local-language explanations, and a feedback channel for extension officers. This would turn the current proof of concept into an operational system. Recent work by Picon et al. [13] demonstrates that crop-conditional CNNs can achieve high performance on mobile devices, suggesting that model distillation of the CLIP-based framework is a viable path forward.

Before practical deployment, several challenges remain. The framework should be validated using independent field datasets collected under diverse environmental conditions, including variations in lighting, background complexity, mixed infections, and crop varieties. In addition, locally adapted reward functions should be developed with agronomists to better reflect regional management practices and disease epidemiology.

Limitations and threats to validity

Despite the strong performance of the proposed framework, several limitations should be acknowledged. First, the image dataset is primarily derived from PlantVillage, whose images are captured under controlled conditions and are generally cleaner than those encountered in real agricultural environments; consequently, model performance may decline under natural lighting, mixed infections, occlusions, and complex backgrounds, as comprehensively discussed by Barbedo [7]. Second, the reinforcement learning reward function is based on expert-defined agronomic rules that are transparent but simplified, whereas real disease management decisions also depend on factors such as disease severity, crop growth stage, seasonal conditions, chemical availability, farmer practices, and local disease pressure. In addition, the reinforcement learning agent is trained using offline case transitions rather than longitudinal observations of disease progression, meaning that the learned policy reflects simulated decision scenarios rather than real temporal disease dynamics. The relatively lower recall for the minority Monitor and Expert actions further suggests the need for class-balanced reward design, uncertainty-aware calibration, and richer examples of ambiguous disease cases. Although SHAP analysis provides valuable insight into the contribution of CLIP embedding dimensions, these latent representations should not be interpreted as explicit biological or pathological features. Furthermore, the current CLIP-based framework has not yet been optimized for deployment on resource-constrained devices such as smartphones or Raspberry Pi-class platforms, highlighting the need for future research on knowledge distillation, quantization, pruning, and latency optimization [13]. Finally, as noted in Section "Contribution of VQA Question Types", the unequal sample sizes across question categories (Table 4) require comparisons of classification accuracy to be interpreted with caution. Although balanced downsampling preserved the overall ranking of model performance, confidence intervals remained wider for categories with fewer samples, indicating greater statistical uncertainty.

How to cite this article:

Delphin M N, Kasambya M K, Kiposo M K, et al. (August 18, 2026) From Disease Recognition to Agronomic Action: Vision-Language Reinforcement Learning for Explainable Crop Monitoring Using PlantVillageVQA. *Cureus J Comput Sci* 3 : es44389-026-00227-3. DOI <https://doi.org/10.7759/s44389-026-00227-3>

Future work

Although the proposed framework provides encouraging results, some limitations should be acknowledged. The experiments were conducted mainly on PlantVillageVQA, which offers a structured and controlled benchmark but does not fully reflect the variability of real farming environments. Field images often contain complex backgrounds, uneven lighting, occlusion, mixed symptoms, and differences in image quality [7]. Future work will therefore extend the evaluation to external datasets such as PlantDoc [8] and to locally collected crop images from Nord-Kivu in order to assess robustness under domain shift. In addition, repeated runs with confidence intervals will be reported to further quantify the stability of the DDQN policy and the baseline models.

Future developments will also focus on improving the safety and practical reliability of the decision-support process. In particular, uncertainty estimation and calibration metrics will be integrated so that the system can recommend expert referral when model confidence is low, rather than relying only on predefined class-action rules. More expressive multimodal fusion strategies, such as gated fusion or lightweight cross-attention, will also be explored to better capture interactions between visual symptoms and textual disease descriptions. Finally, the reward rules and class-action mappings will be reviewed with local agronomists and plant-health experts to ensure that the recommendations remain scientifically valid, locally appropriate, and useful for smallholder farmers.

Beyond technical improvements, future studies should conduct prospective field trials involving agricultural extension officers and farmers to evaluate usability, trust, decision quality, and the practical impact of the recommendations on crop management. Such evaluations will provide stronger evidence of the framework's effectiveness under real agricultural conditions.

Conclusions

This study presented a vision-language reinforcement-learning framework for explainable crop monitoring and agronomic decision support. By fusing CLIP image embeddings and PlantVillageVQA text embeddings, the framework converts multimodal evidence into a four-action policy: monitor, treat, isolate, or refer to an expert. The DDQN policy achieved 94.4% accuracy and 0.861 macro-F1-score, while fusion baselines showed that textual VQA context adds substantial value over image-only representations.

The main contribution is not a claim of universal superiority over all classifiers. Instead, the manuscript contributes a reproducible evidence-to-action design for agricultural AI: multimodal representation, expert-defined action reward, policy learning, ablation of question types, and action-level explanations. The findings suggest that causal and descriptive language can strengthen crop disease decision support, particularly when the objective is to guide intervention rather than merely name a disease. Although the present work represents a proof-of-concept developed under controlled experimental conditions, it demonstrates that integrating multimodal vision-language learning, reinforcement learning, and explainable artificial intelligence is a promising direction for moving beyond disease recognition toward transparent, action-oriented agricultural decision support. Future validation using real-world field datasets, locally adapted reward functions, and lightweight deployment strategies will be essential for translating this framework into a reliable tool for precision agriculture and smallholder farming.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Morisho N. Delphin, Moïse K. Kasambya, Marc K. Kiposo, Héritier K. Kaheraya, Francis N. Mukumbi, Daniel M. Kalenga, Dieudonné E. Odjedra, Wasingya K. Maunga, Furaha K. Vayihonga

Acquisition, analysis, or interpretation of data: Morisho N. Delphin, Moïse K. Kasambya, Marc K. Kiposo, Héritier K. Kaheraya, Francis N. Mukumbi, Daniel M. Kalenga, Dieudonné E. Odjedra, Wasingya K. Maunga, Furaha K. Vayihonga

How to cite this article:

Delphin M N, Kasambya M K, Kiposo M K, et al. (August 18, 2026) From Disease Recognition to Agronomic Action: Vision-Language Reinforcement Learning for Explainable Crop Monitoring Using PlantVillageVQA. *Cureus J Comput Sci* 3 : es44389-026-00227-3. DOI <https://doi.org/10.7759/s44389-026-00227-3>

Drafting of the manuscript: Morisho N. Delphin, Moïse K. Kasambya, Marc K. Kiposo, Héritier K. Kaheraya, Francis N. Mukumbi, Daniel M. Kalenga, Dieudonné E. Odjedra, Wasingya K. Maunga, Furaha K. Vayihonga

Critical review of the manuscript for important intellectual content: Morisho N. Delphin, Moïse K. Kasambya, Marc K. Kiposo, Héritier K. Kaheraya, Francis N. Mukumbi, Daniel M. Kalenga, Dieudonné E. Odjedra, Wasingya K. Maunga, Furaha K. Vayihonga

Supervision: Morisho N. Delphin

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The data that support the findings of this study are openly available in Kaggle at <https://www.kaggle.com/datasets/ahmedmeliane/plantvillagevqa2>

References

1. Savary S, Willcoquet L, Pethybridge SJ, Esker P, McRoberts N, Nelson A: [The global burden of pathogens and pests on major food crops](#). Nature Ecology & Evolution. 2019, 3:430-439. [10.1038/s41559-018-0793-y](#)
2. Mohanty SP, Hughes DP, Salathé M: [Using deep learning for image-based plant disease detection](#). Frontiers in Plant Science. 2016, 7:1419. [10.3389/fpls.2016.01419](#)
3. Ferentinos KP: [Deep learning models for plant disease detection and diagnosis](#). Computers and Electronics in Agriculture. 2018, 145:311-318. [10.1016/j.compag.2018.01.009](#)
4. Mpia HN, Mburu LW, Mwendia SN: [Applying data mining in graduates' employability: a systematic literature review](#). International Journal of Engineering Pedagogy (IJEP). 2023, 13:86-108. [10.3991/ijep.v13i2.33643](#)
5. Radford A, Kim JW, Hallacy C, et al.: [Learning transferable visual models from natural language supervision \[PREPRINT\]](#). arXiv. 2021, [10.48550/arXiv.2103.00020](#)
6. Mpia HN, Kavalami MM, Lusenge GK, Ushindi KP, Kambale Kyalengekanja DD, Ciswa OM: [Lightweight deep learning system for infant-cry recognition with real-time notification in resource-constrained environments](#). Machine Learning with Applications. 2025, 22:100791. [10.1016/j.mlwa.2025.100791](#)
7. Barbedo JGA: [Factors influencing the use of deep learning for plant disease recognition](#). Biosystems Engineering. 2018, 172:84-91. [10.1016/j.biosystemseng.2018.05.013](#)
8. Singh D, Jain N, Jain P, et al.: [PlantDoc: a dataset for visual plant disease detection](#). CoDS COMAD 2020: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD. 2020, 249-253. [10.1145/3371158.3371196](#)
9. Alibabaei K, Gaspar PD, Lima TM, et al.: [A review of the challenges of using deep learning algorithms to support decision-making in agricultural activities](#). Remote Sensing. 2022, 14:638. [10.3390/rs14030638](#)
10. Yue R, He L, Ye R, et al.: [Distributed crop disease detection with deep Q-network agent decision-making in a compressed state space](#). Artificial Intelligence in Agriculture. 2026, 16:1145-1163. [10.1016/j.aiia.2026.02.006](#)
11. Lundberg SM, Lee SI: [A unified approach to interpreting model predictions \[PREPRINT\]](#). arXiv. 2017, [10.48550/arXiv.1705.07874](#)
12. Hu G, Yang Y, Yi D, et al.: [When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition \[PREPRINT\]](#). arXiv. 2015, [10.48550/arXiv.1504.02351](#)

How to cite this article:

Delphin M N, Kasambya M K, Kiposo M K, et al. (August 18, 2026) From Disease Recognition to Agronomic Action: Vision-Language Reinforcement Learning for Explainable Crop Monitoring Using PlantVillageVQA. Cureus J Comput Sci 3 : es44389-026-00227-3. DOI <https://doi.org/10.7759/s44389-026-00227-3>

13. Picon A, Seitz M, Alvarez-Gila A, Mohnke P, Ortiz-Barredo A, Echazarra J: [Crop conditional convolutional neural networks for massive multi-crop plant disease classification over cell phone acquired images taken on real field conditions](#). Computers and Electronics in Agriculture. 2019, 167:105093. [10.1016/j.compag.2019.105093](#)
14. Van Hasselt H, Guez A, Silver D: [Deep reinforcement learning with double Q-learning](#). Proceedings of the AAAI Conference on Artificial Intelligence. 2016, 30:2094-2100. [10.1609/aaai.v30i1.10295](#)
15. Araya T, Ochsner TE, Mnkeni PNS, Hounkpatin KOL, Amelung W: [Challenges and constraints of conservation agriculture adoption in smallholder farms in sub-Saharan Africa: A review](#). International Soil and Water Conservation Research. 2024, 12:828-843. [10.1016/j.iswcr.2024.03.001](#)

How to cite this article:

Delphin M N, Kasambya M K, Kiposo M K, et al. (August 18, 2026) From Disease Recognition to Agronomic Action: Vision-Language Reinforcement Learning for Explainable Crop Monitoring Using PlantVillageVQA. Cureus J Comput Sci 3 : es44389-026-00227-3. DOI <https://doi.org/10.7759/s44389-026-00227-3>