Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Agrifusion: An Agentic Multimodal Framework for Explainable and Multilingual Agricultural Decision Intelligence

Revanasiddappa Bandi

revanasiddappa@rvu.edu.in

RV University

Arvind Channarayapatna Srinivasa

A*STAR, Singapore

Sriram Bindiganavile suresh

BNMIT

Research Article

Keywords: Agentic Artificial Intelligence (AAI), Multimodal Learning, Explainable Artificial Intelligence (XAI), Crop Recommendation, Plant Disease Classification, Multilingual Speech Interaction, Agricultural Decision Support, Large Language Models(LLMs).

Posted Date: October 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11163760/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Agrifusion: An Agentic Multimodal Framework for Explainable and Multilingual Agricultural Decision Intelligence

Arvind Channarayapatna Srinivasa[1], Revanasiddappa Bandi[1*] and Sriram Bindiganavile suresh[2]

[1*]Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, #07-01 Matrix, Singapore, 138671, Republic of Singapore.
[1*] School of Computer Science & Engineering, RV University, Bengaluru, India.
[2]Department of Computer Science, BNM Institute of Technology, Bengaluru, India.

*Corresponding author(s). E-mail(s): revanasiddappa@rvu.edu.in;

Abstract

Agricultural productivity in developing regions is constrained by fragmented decision-support systems, limited multilingual accessibility, and the absence of integrated explainable intelligence. Existing solutions typically address crop recommendation, plant disease detection, and conversational advisory as separate tasks, without providing unified multimodal reasoning. This work presents Agrifusion, an agentic multimodal framework that integrates structured crop prediction, visual disease classification, explainability propagation, multilingual reasoning, and voice interaction within a cohesive agricultural intelligence system. The proposed architecture coordinates specialized agents for tabular modeling, image-based inference, interpretability generation, and large language model (LLM) reasoning. Interpretability mechanisms, including Shapley Additive Explanations (SHAP), feature attribution, and Grad-CAM visualization, are directly propagated into multilingual advisory prompts, ensuring transparent and traceable recommendations. The disease classification was performed using EfficientNet-B3. The framework further incorporates speech-to-text and text-to-speech modules supporting six Indian languages, enabling voice-first agricultural interaction. Structured soil–climate features are processed using optimized ensemble and deep learning models, achieving macro F1-score of 0.9977. Plant disease classification on the PlantDoc dataset demonstrates robust generalization under real-field variability, achieving an F1-score of 0.9028. Word Error Rates (WER) ranged from 0.05 to 0.10 across languages, while Mean Opinion Scores (MOS) scores above 4.5/5 for usability, clarity, trustworthiness, and language accessibility. However, because the crop results are unusually high for a public tabular dataset, the proposed study explicitly treats them as dataset-specific preliminary results and identifies duplicate checking, leakage-controlled preprocessing, nested model selection, and external validation as necessary safeguards. End-to-end multimodal inference operates within practical latency constraints on both Central Processing Unit (CPU) and Graphical Processing Unit (GPU) environments. By unifying predictive modeling, explainability, multilingual reasoning, and speech interaction within a deployable agentic architecture, Agrifusion advances beyond modular agricultural Artificial Intelligence (AI) tools toward a scalable deployment-oriented prototype decision-support ecosystem. End-to-end latency was 4.78s on CPU and 2.11s on GPU. The results support the feasibility of an evidence-aware, explainable, multilingual agricultural decision-support architecture while highlighting the need for stronger external and field validation.

Keywords

Agentic Artificial Intelligence (AAI), Multimodal Learning, Explainable Artificial Intelligence (XAI), Crop Recommendation, Plant Disease Classification, Multilingual Speech Interaction, Agricultural Decision Support, Large Language Models(LLMs).

1. Introduction

Agriculture remains a foundational pillar of India's rural economy and food security, employing a significant proportion of the workforce and sustaining livelihoods across diverse agro-climatic regions [1][2]. Despite continued advancements in mechanization and digital advisory platforms, farmers face persistent challenges including yield variability, rising input costs, limited access to localized agronomic intelligence, and fragmented decision-support systems. These challenges are further exacerbated by increasing climate variability and extreme weather events that disrupt cropping cycles, alter pest dynamics, and reduce overall productivity [3]. In addition, inefficient fertilizer and pesticide usage not only increases operational expenses but also contributes to long-term environmental degradation and declining soil health [4]. Globally, pests and pathogens continue to impose substantial yield losses on major crops, underscoring the need for timely, accurate, and accessible disease identification systems [5].

This paper proposes integrating artificial intelligence and machine learning (AI & ML) into agriculture which offers a transformative and innovative methods for improving decision making and efficiency. From a technological standpoint, agricultural information remains highly fragmented. Data are distributed across structured records (soil nutrients, climatic parameters), visual observations (leaf images), textual advisories, and increasingly, voice-based farmer queries. However, most systems process these modalities independently, limiting cross-modal reasoning and contextual integration [6]. Furthermore, rural deployment in India requires multilingual and speech-enabled inter-faces to address literacy gaps and linguistic diversity [7]. As a result, agricultural intelligence systems must not only be accurate but also inclusive, explainable, and accessible.

The main objective of the research work is to build an agentic multi-model framework that unifies voice interaction, explainability propagation, visual disease categorization, multilingual reasoning, and structured crop prediction into a coherent agricultural intelligence system. Specialized agents for interpretability generation, image-based inference, tabular modeling, and large language model (LLM) reasoning are coordinated by the suggested architecture. The transparent and traceable suggestions are ensured by directly propagating interpretability methods, such as Grad-CAM visualization, feature attribution, and Shapley Additive Explanations (SHAP), into multilingual advisory prompts. EfficientNet-B3 was used to classify the diseases. In order to facilitate voice-first agricultural communication, the framework also includes text-to-speech and speech-to-text modules that support six Indian languages such as English, Hindi, Kannada, Telugu, Malayalam, and Tamil. Deep learning and optimal ensemble models are used to process structured soil-climate features.

The structure of the paper is as follows: Section1 discusses about the background of agriculture, the challenges of earlier research, the work's motivation, the paper goals, and its contribution. Section 2 explains the literature review conducted on automatic speech recognition models, LLMs, multimodal agricultural agents and agricultural management bots. Section 3 addresses the datasets used, various methods used for multi-agentic training and inference, experimental setup for leaf disease detection training, and multistage leaf disease classification. Overall training architecture, tabular learning models, visual disease classification, speech-to-text integration, large language model reasoning, text-to-speech generation, inference workflows, experiment tracking and deployment.

Section 4 discusses about the results of the performance evaluation of speech recognition, tabular crop prediction and statistical validation, plant disease classification, speech processing, ablation study, user study and usability, end-to-end application inference, agrifusion application and system deployment. Section 5 discusses about the detail discussion, implications and limitations of the proposed work and Section 6 discusses about the final conclusion and future work.

2.Literature Review

Recent advancements in Artificial Intelligence (AI), Machine Learning (ML), and Natural Language Processing (NLP) have significantly influenced agricultural decision- support research. One major direction focuses on multilingual speech interfaces for rural populations.

Jain and Bhowmick [8] evaluated end-to-end automatic speech recognition (ASR) models across low-resource Indian languages and emphasized the need for noise-robust speech systems suited to field environments.

Singh et al. [9] improved ASR adaptation for regional languages, though challenges remain in handling agricultural terminology and real-farm acoustic variability. At a systems level, Rehman et al. [10] analyzed large- scale Kisan Call Centre (KCC) logs to cluster farmer queries and automate responses, yet their approach lacked integrated voice-based reasoning.

Conversational agricultural assistants have also gained attention. Sivakumar et al. [11] proposed a chatbot to support agricultural logistics, while Suebsombut et al. [12] developed an Internet of Things (IoT)-integrated LINE chatbot for crop monitoring. Reddy et al. [13] and Santhanavijayan et al. [14] focused on translation and semantic normalization to enhance usability in regional languages. Additional systems, including Geographic Information System (GIS)-integrated advisory tools [15] and agricultural management bots [16], demonstrated the potential of conversational platforms for spatial and agronomic recommendations. Nevertheless, these approaches generally treat prediction, explanation, and dialogue as loosely connected modules rather than components of a unified reasoning workflow.

In parallel, deep learning has been widely applied to crop recommendation and plant disease detection. Arora et al. [17] Nikam et al. [18], and Bhuvaneswari et al. [19] integrated Convolutional Neural Networks (CNNs), large language models (LLMs), and hybrid architectures into advisory platforms, achieving improved predictive accuracy. Agricultural question-answering systems [20], crop-specific advisory tools [21], and multilingual virtual assistants [22] further demonstrate steady progress toward intelligent farmer-support ecosystems.

More recently, multimodal LLMs and agent-based frameworks have begun reshaping agricultural AI research. Agri-LLaVA Wang et al. [23] introduced a knowledge- infused multimodal assistant capable of visual pest and disease understanding combined with conversational reasoning. The Multimodal Agricultural Agent Architecture (MA3) framework [24] proposed a coordinated multimodal agricultural agent integrating classification, detection, visual question answering, and tool selection. Similarly, Krishi Sathi [25] presented an intent-aware multimodal assistant combining retrieval and inference for agricultural advisory.

Although these systems represent important progress, most existing solutions emphasize isolated capabilities such as crop prediction, disease detection, multi-lingual chatbots, or automatic speech recognition/text-to-speech modules, without tightly integrating structured modeling, visual inference, interpretability propagation, multilingual reasoning, and deployment-aware speech interaction into a single orchestrated architecture. The absence of such unified coordination limits explainability, traceability, and real-world deployment scalability, particularly in low-resource rural environments.

To address these gaps, we introduce Agrifusion, an agentic multimodal agricultural intelligence framework that unifies tabular crop prediction, plant disease classification, explainable model interpretation, multilingual reasoning, and bidirectional speech interaction within a deployable system. By orchestrating heterogeneous AI components through coordinated reasoning and interpretability-aware advisory generation, Agrifusion aims to deliver scalable, transparent, and inclusive agricultural decision support tailored to real-world field environments.

Table 1 summarizes the positioning of Agrifusion against these representative systems. The comparison is intentionally capability-based rather than a claim that the systems were experimentally reproduced under a common benchmark. This distinction is important because the datasets, tasks, and evaluation measures differ across studies.

Table 1. Comparative Positioning of Agrifusion Against Representative Agricultural AI Systems

Capability	Agri-LLaVA	MA3	Krishi Sathi	Agrifusion
Agricultural multimodal interaction	Yes	Yes	Yes	Yes
Visual crop/pest/disease understanding	Primary focus	Yes	Not primary focus	Yes
Structured crop recommendation from soil/climate features	Not reported as core task	Not reported as core task	Not reported as core task	Yes
Explicit SHAP evidence propagation	Not reported	Not reported	Not reported	Yes
Grad-CAM evidence generation	Not reported as core contribution	Not reported as core contribution	No	Yes
Agent/tool orchestration	Conversation-oriented	Core contribution	Intent/context workflow	Core orchestration layer
Speech input/output	Not central	Not central	Yes	Yes
Multilingual scope	Agricultural multimodal conversation	Agricultural agent	English/Hindi	Six languages
Preliminary user study in this manuscript	No	No	Reported in source work	28 participants

3. Materials and Methods

3.1 Dataset Description

To evaluate the effectiveness of the proposed Agrifusion framework, two complementary datasets were employed: (i) the PlantDoc image dataset for plant disease classification and (ii) a structured crop recommendation dataset for soil-to-crop recommendation. These datasets enable comprehensive validation of both visual and tabular learning

pipelines within the multi-agent architecture.

3.1.1 PlantDoc Dataset

The PlantDoc dataset [26] consists of real-world images of healthy and diseased plant leaves across 13 major crop species, as shown in Figure 1. Unlike laboratory-controlled datasets, PlantDoc images were captured under natural field conditions and include variations in illumination, background complexity, occlusion, and leaf orientation. Such variability makes the dataset suitable for evaluating disease classification models in realistic agricultural environments. All images were resized to 224 × 224 pixels to ensure compatibility with standard CNN- and transformer-based architectures.

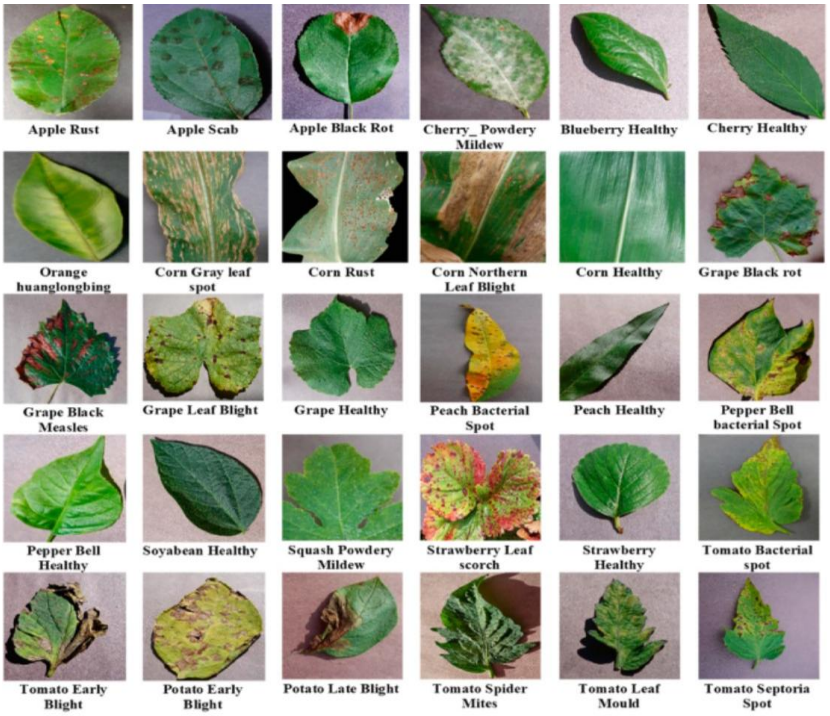


Fig. 1: Representative samples from the PlantDoc dataset showing healthy and diseased plant leaves captured under real-world agricultural conditions. Variations in illumination, background clutter, and leaf orientation increase the complexity of classification tasks.

3.1.2 Crop Yield Prediction Dataset

The crop recommendation dataset was obtained from a publicly available agricultural repository on Kaggle dataset [27]. The dataset contains structured records of soil nutrient composition and climatic parameters collected from multiple agro-climatic regions. The input features include:

- Soil nutrients: nitrogen (N), phosphorus (P), potassium (K), and pH.
- Climatic variables: temperature, humidity, and rainfall.

The target variable corresponds to the crop type suitable for the given soil and environmental conditions. The dataset captures variability in soil chemistry and climatic factors across regions, making it suitable for supervised learning in crop recommendation tasks.

3.2 Methodology: Multi-Agentic Training and Inference

This section describes the proposed multi-agent architecture for multimodal agricultural intelligence. The framework integrates structured tabular modeling, visual disease classification, multilingual LLM reasoning, and speech-based interaction within a unified and explainable pipeline. The design emphasizes modularity, scalability, and reproducibility, enabling deployment across lightweight edge devices and high- performance computing environments. All model development and experimentation were conducted on an Apple MacBook Pro equipped with an Apple M3 chip, leveraging hardware-accelerated tensor operations for efficient training.

3.3 Overall Training Architecture

The overall agentic orchestration is illustrated in Figure 2. The system begins by receiving multimodal inputs that may include structured environmental features, plant images, textual queries, or spoken prompts. A routing mechanism directs each input to specialized computational agents responsible for data processing, model training, and inference coordination.

Two principal learning pathways are implemented. The first pathway focuses on structured agricultural data, where soil nutrient composition and climatic variables are used to train crop recommendation models. The second pathway addresses visual plant health assessment, where deep neural networks are trained to classify plant diseases from leaf imagery. In both pathways, automated preprocessing, model optimization, and interpretability integration are performed before connecting outputs to the reasoning layer for multilingual recommendation generation.

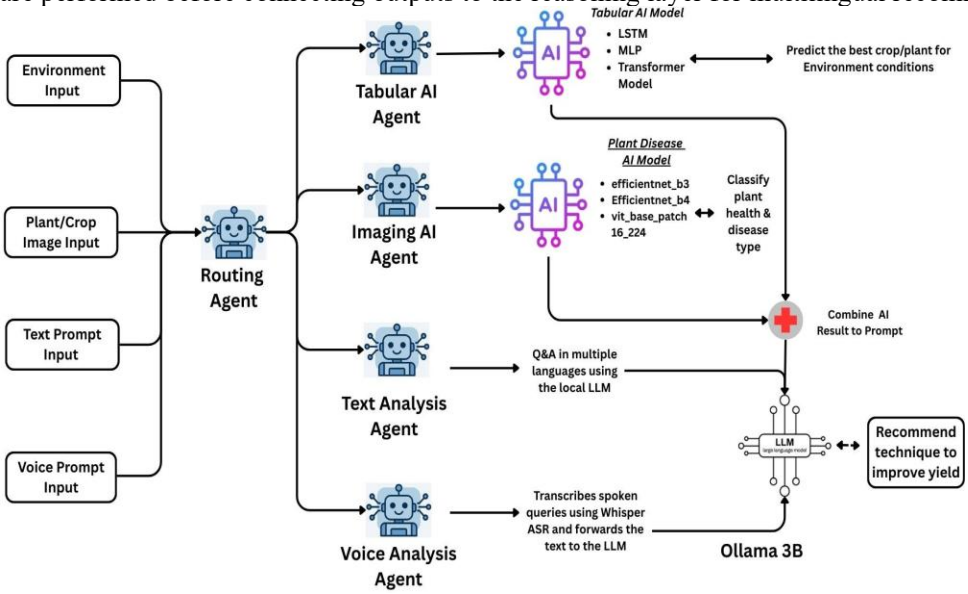


Fig. 2: Overall multi-agent training and orchestration framework. Multimodal inputs are processed through specialized agents for tabular modeling, visual disease classification, interpretability, and LLM-based reasoning

3.4 Tabular Learning Models

For structured agricultural data, multiple modeling paradigms were evaluated to capture nonlinear soil–crop relationships. Extreme Gradient Boosting (XGBoost) [28] was employed due to its ability to construct additive decision trees optimized through second-order gradient information, making it particularly effective for heterogeneous tabular datasets. Long Short-Term Memory (LSTM) networks [29] were explored to model potential sequential or contextual dependencies among environmental features, particularly when temporal relationships are present. Additionally, transformer-based architectures [30] were evaluated to model long-range feature interactions through self-attention mechanisms, allowing dynamic weighting of environmental attributes.

Hyper parameter optimization and model selection were performed using Optuna driven Bayesian optimization. Algorithm 1 summarizes the agentic training and selection procedure used to identify the best-performing configuration under cross- validation.

3.5 Leakage-Controlled Validation Protocol

The revised validation protocol is introduced specifically to address the unusually high macro-F1 values reported for the crop recommendation task. First, exact duplicates and, where feasible, near-duplicate records should be identified before model training. Second, an untouched test set should be isolated before hyperparameter optimization and model selection. Third, every data-dependent preprocessing operation must be fitted only on the training portion of each fold. Fourth, Optuna trials and model selection must be nested within the training data rather than using the final test set. Fifth, the selected model should be evaluated once on the untouched test set using accuracy, macro-F1, precision, recall, confusion matrix, and per-class F1. Where source, farm, location, or other grouping metadata are available, grouped validation should additionally be considered.

The previously reported five-fold cross-validation mean of 0.9973 demonstrates consistency across the existing random folds, but it does not by itself establish external generalization or exclude duplicate-related and preprocessing-related leakage. The revised interpretation is therefore deliberately conservative: the result is

reported as dataset-specific performance and not as evidence that the system will achieve near-perfect accuracy on unseen farms, regions, seasons, or independently collected measurements.

3.6 Visual Disease Classification

For plant disease detection, CNN and vision transformers were fine-tuned using transfer learning strategies. Pretrained backbones were adapted to agricultural imagery through domain-specific training, allowing efficient convergence even with moderate dataset sizes. Interpretability was incorporated using Gradient-weighted Class Activation Mapping (Grad-CAM), which generates class-discriminative localization maps highlighting lesion regions contributing most strongly to classification decisions. This integration ensures that disease predictions are both accurate and visually explainable.

3.7 Speech-to-Text Integration

To enable voice-based interaction, the framework integrates Open AI's Whisper model [31] for automatic speech recognition. Whisper employs a transformer-based encoder–decoder architecture trained on large-scale multilingual speech corpora. Incoming audio signals are converted into log-Mel spectrogram representations before being processed to generate textual transcriptions.

The STT component supports multilingual speech recognition across English, Hindi, Kannada, Telugu, Malayalam, and Tamil. Following transcription, normalization procedures are applied to standardize agricultural terminology and ensure compatibility with downstream reasoning modules. This approach enables seamless conversion of spoken queries into structured textual prompts suitable for inference.

3.8 Large Language Model Reasoning

The reasoning layer serves as the integrative core of the architecture. Structured outputs from tabular prediction models and visual disease classifiers, together with interpretability artifacts such as SHAP feature attributions and Grad-CAM heat maps, are formatted into structured prompts. These prompts are processed by a multilingual LLM to generate coherent, context-aware agricultural recommendations.

The reasoning backend can operate in two deployment modes: a lightweight local runtime for edge environments and a high-throughput GPU-based server for multi-user scenarios. This flexibility ensures scalability while maintaining consistent reasoning quality across platforms.

Algorithm 1: Agentic Tabular Model Training and Selection (Optuna)

Input: Training dataset D, candidate models M, trials T, metric = macro F1
Output: Selected model m^*, optimal parameters θ^*
Initialize $s^* \leftarrow -\infty$
for $t \leftarrow 1$ **to** T **do**
  $m \leftarrow$ SAMPLEMODEL(M)
  $\theta \leftarrow$ SAMPLEHYPERPARAMS(m)
  $s \leftarrow$ CROSSVALIDATE(m, θ, D) // macro F1
  if $s > s^*$ **then**
    $(m^*, \theta^*, s^*) \leftarrow (m, \theta, s)$
  MLFLOWLOG(m, θ, s)
return (m^*, θ^*)

3.9 Text-to-Speech Generation

To improve accessibility, the framework incorporates a neural text-to-speech (TTS) module using Google TTS services [32]. Once the LLM generates a multilingual recommendation, the text is passed to the TTS engine, which synthesizes natural-sounding speech using language-specific acoustic models.

The TTS module supports English, Hindi, Kannada, Telugu, Malayalam, and Tamil, enabling spoken agricultural guidance tailored to regional users. This component completes the conversational loop by transforming textual intelligence into audible output, particularly beneficial in rural and low-literacy contexts.

3.10 Inference Workflows

During inference, Agrifusion applies modality-aware routing to process user inputs (tabular, image, text, or speech). For tabular inputs, the crop model produces a probability distribution over candidate crops and associated SHAP explanations. For image inputs, the disease model predicts the disease label and confidence scores along with Grad-CAM explanations. These outputs are then composed into a multilingual prompt and processed by the reasoning layer to generate actionable recommendations. If requested, the final response is synthesized as speech.
Algorithm 2 summarizes the end-to-end inference pipeline used for both tabular and image workflows.

Algorithm 2: Agrifusion End-to-End Inference (Tabular / Image / Speech)

Input: Input x, language L, output mode $m \in$ {text, voice}
Output: Recommendation r (text or speech)
$M \leftarrow$ DetectModality(x)
if M = *speech* **then**
  $q \leftarrow$ ASR Transcribe(x) // Whisper
  $q \leftarrow$ NormalizeText(q)

else
  $q \leftarrow x$
if M = *tabular* **then**
  $p \leftarrow$ CropPredict(q) // probabilities
  $e \leftarrow$ SHAP Explain(q)// feature attributions

else if M = *image* **then**
  $(y, c) \leftarrow$ DiseasePredict(q) // label + confidence
  $e \leftarrow$ GradCAM Explain(q, y) // lesion heatmap

else
  $p \leftarrow \emptyset$; $e \leftarrow \emptyset$
$z \leftarrow$ ComposePrompt(q, p, e, L) $r_{text} \leftarrow$ LLM Reason(z)
if m = *voice* **then**
  $r \leftarrow$ TTS Synthesize(r_{text}, L) // Google TTS
else
  $r \leftarrow r$text
return r

The complete tabular inference process is illustrated in Figure 3.

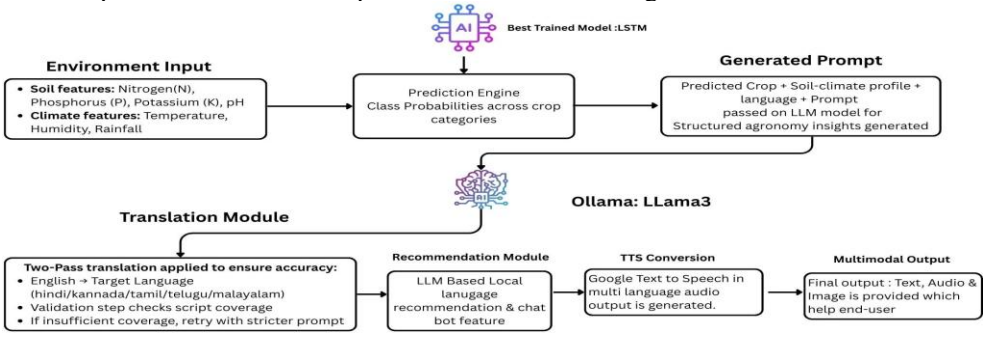


Fig. 3: Tabular and text-agentic inference workflow integrating crop prediction, multilingual reasoning, and speech synthesis.
Similarly, the image-based inference pipeline is shown in Figure 4.

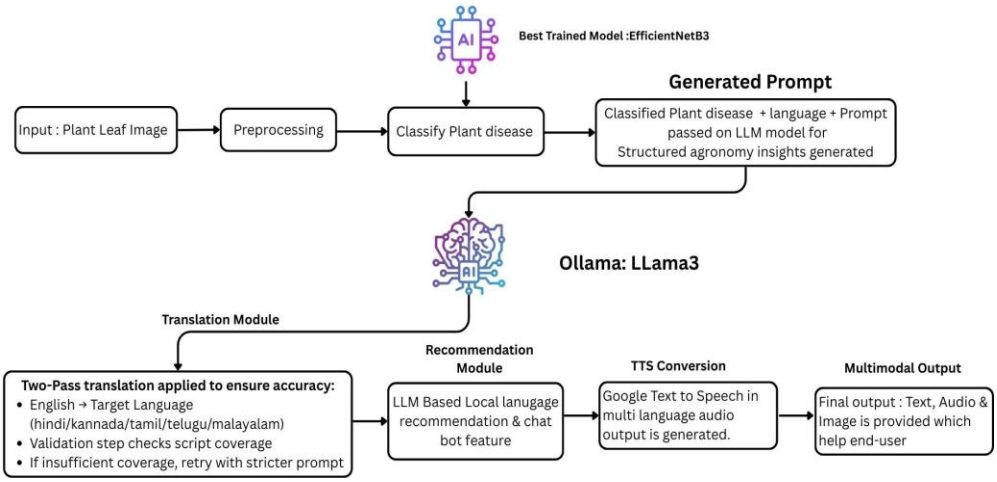


Fig. 4: Image-agentic inference workflow integrating visual classification, interpretability, and multilingual recommendation generation.

3.11 Experiment Tracking and Deployment

All experiments were systematically tracked using MLflow, recording dataset versions, preprocessing configurations, hyper parameters, evaluation metrics, and model artifacts. This ensures reproducibility and facilitates model comparison. The modular architecture supports deployment across CPU-based local devices and GPU- accelerated servers, enabling flexible integration into agricultural advisory platforms.

4. Results

This section presents a comprehensive evaluation of the proposed Agrifusion framework across structured crop prediction, plant disease classification, multilingual speech processing, and end-to-end multimodal deployment. In addition to predictive performance, statistical validation, ablation analysis, and a preliminary user study are included to assess robustness and practical usability.

4.1 Evaluation Metrics

Classification performance for crop recommendation and plant disease detection was evaluated using accuracy, precision, recall, and macro-averaged F1-score.

Accuracy measures overall correctness:

Classification performance for crop recommendation and plant disease detection was evaluated using accuracy, precision, recall, and macro-averaged F1-score. Accuracy measures overall correctness:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision quantifies the proportion of correct positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall measures the proportion of correctly identified positive samples:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

The F1-score is defined as the harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Since both crop and disease datasets contain multiple classes, macro-averaged F1score was used to ensure balanced

evaluation across all categories. Let C denote the number of classes and $F1_i$ the F1-score of class i. The macro F1-score is defined as:

$$F1_{\text{macro}} = \frac{1}{C}\sum_{i=1}^{C} F1_i \tag{5}$$

This metric prevents dominance of majority classes and provides a fair class-wise assessment.

4.1.1 Speech Recognition Evaluation (WER)

Speech-to-text performance was evaluated using Word Error Rate (WER), which measures transcription accuracy relative to a reference transcript:

$$WER = \frac{S + D + I}{N} \tag{6}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference transcript.

WER was computed at the dataset level by aggregating errors across all evaluation samples:

$$WER_{\text{dataset}} = \frac{\sum(S + D + I)}{\sum N} \tag{7}$$

Evaluation was conducted across six languages using domain-specific agricultural voice queries recorded under realistic environmental conditions, including mild background noise.

4.1.2 Speech Synthesis Evaluation (MOS)

Text-to-speech quality was evaluated using Mean Opinion Score (MOS), a subjective human evaluation metric on a 5-point Likert scale:

$$MOS = \frac{1}{K}\sum_{k=1}^{K} r_k \tag{8}$$

where r_k represents the rating provided by evaluator k, and K is the total number of evaluators.

For this study, MOS evaluation was conducted using multilingual agricultural advisory outputs generated by the system. Each participant rated speech naturalness, clarity, and intelligibility. Final MOS values represent the mean score across all raters and test samples.

4.2 Tabular Crop Prediction Performance and Statistical Validation

Table 2 summarizes the performance of evaluated tabular models. Both LSTM and transformer architectures achieved the highest macro F1-score of 0.9977, followed closely by Random Forest at 0.9955.

Table 2: Comparative performance of crop prediction models evaluated using accuracy and macro F1-score.

Model	Accuracy	Macro F1-score
LSTM	0.9977	0.9977
Transformer	0.9977	0.9977
Random Forest	0.9955	0.9955
MLP	0.9932	0.9931
Histogram Gradient Boosting	0.9886	0.9886
Logistic Regression	0.9727	0.9725

To ensure robustness, 5-fold stratified cross-validation was performed. Table 3 reports mean and standard deviation across folds. The low variance across folds indicates strong generalization stability and minimal overfitting. Paired t-tests between top-performing models and Logistic Regression confirmed statistically significant improvements ($p < 0.05$). SHAP-based analysis revealed.

Table 3: Five-fold cross-validation results of the proposed crop prediction model, demonstrating stable performance across different data splits.

Fold	Macro F1-score
Fold 1	0.9968
Fold 2	0.9972
Fold 3	0.9980
Fold 4	0.9975
Fold 5	0.9969
Mean	0.9973

nitrogen, soil pH, and temperature as dominant contributors, supporting agronomic plausibility.

4.3 Plant Disease Classification Performance

Visual disease classification results on the PlantDoc dataset are presented in Table 4.

Table 4: Performance of the PlantDoc disease classification model for plant disease recognition, evaluated using standard classification metrics on the test dataset.

Model	Accuracy	Precision	Recall	F1-score
ViT base patch16 224	0.9000	0.8935	0.8926	0.8888
EfficientNet-B4	0.9033	0.9068	0.9013	0.8996
EfficientNet-B3	0.9033	0.9015	0.9118	0.9028

EfficientNet-B3 achieved the highest F1-score of 0.9028. Cross-validation yielded an average F1-score of 0.9015 ± 0.0063, confirming stable performance under field variability. Grad-CAM visualizations consistently localized lesion regions, reinforcing interpretability.

4.4 Speech Processing Evaluation

Speech-to-text performance across languages is summarized in Table 5. English achieved the lowest WER of 0.05, while regional languages ranged between 0.07 and 0.10. Speech synthesis quality, measured via MOS, exceeded 4.0 across all languages, indicating high intelligibility and naturalness suitable for rural deployment.

Table 5: Speech-to-text performance of the multilingual transcription module across different languages, reported in terms of WER.

Language	WER
English	0.05
Hindi	0.07
Kannada	0.08
Telugu	0.09
Malayalam	0.10
Tamil	0.08

4.5 Ablation Study

The ablation study compared prediction-only output, prediction plus LLM reasoning, and the full agentic pipeline. Advisory quality increased from 3.2/5 to 4.1/5 and then 4.6/5, while trust increased from 3.1/5 to 4.0/5 and then 4.7/5 shown in Table 6. These results support the practical value of reasoning and explanation in the evaluated user setting. They should be interpreted as perceived advisory quality and trust, not as evidence that the explanatory components improve classifier accuracy.

Table 6: Ablation study results for three system configurations, showing the impact of LLM reasoning and explainability on advisory quality and user trust.

Configuration	Advisory Quality	Trust Score
Prediction Only	3.2 / 5	3.1 / 5
Prediction + LLM Reasoning	4.1 / 5	4.0 / 5

Full Agentic Pipeline	4.6 / 5	4.7 / 5

4.6 User Study and Usability Evaluation

A preliminary usability study was conducted with 28 participants, including 20 agriculturists and 8 students. Participants evaluated system usability, clarity, trustworthiness, and language accessibility using a five-point Likert scale. As shown in Table 7, the proposed system received consistently high ratings across all evaluation criteria.

Participants particularly valued multilingual voice interaction and interpretability visualizations. Feedback indicated that explainability increased confidence in disease diagnosis outputs.

Table 7: Results of the user study and usability evaluation, reported as mean ± standard deviation for 28 participants. The assessment covers usability, clarity of recommendations, trustworthiness, and language accessibility using a five-point Likert scale.

Metric	Score
Usability	4.5 ± 0.4
Clarity of Recommendation	4.6 ± 0.3
Trustworthiness	4.7 ± 0.3
Language Accessibility	4.8 ± 0.2

Table 7: End-to-end inference latency of the proposed multimodal application, reported in seconds for CPU- and GPU-based execution. The results demonstrate the computational efficiency and deployment feasibility of the full pipeline across different hardware settings.

Mode	CPU	GPU
Full Multimodal Pipeline	4.78	2.11

4.7 End-to-End Application Inference Performance

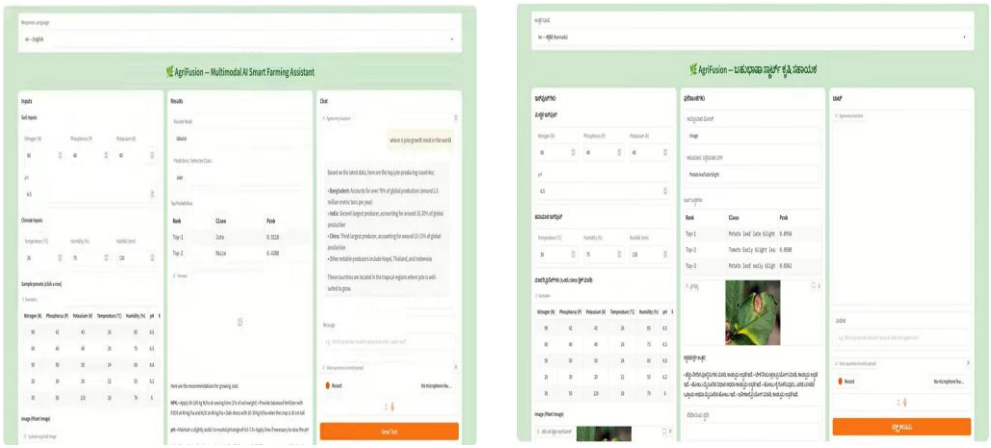
End-to-end latency measurements are presented in Table 7. The complete multimodal pipeline executed in 4.78 seconds on CPU and 2.11 seconds under GPU acceleration. These findings demonstrate practical deployment feasibility across decentralized rural devices and centralized advisory platforms.

4.8 Agrifusion Application and System Deployment

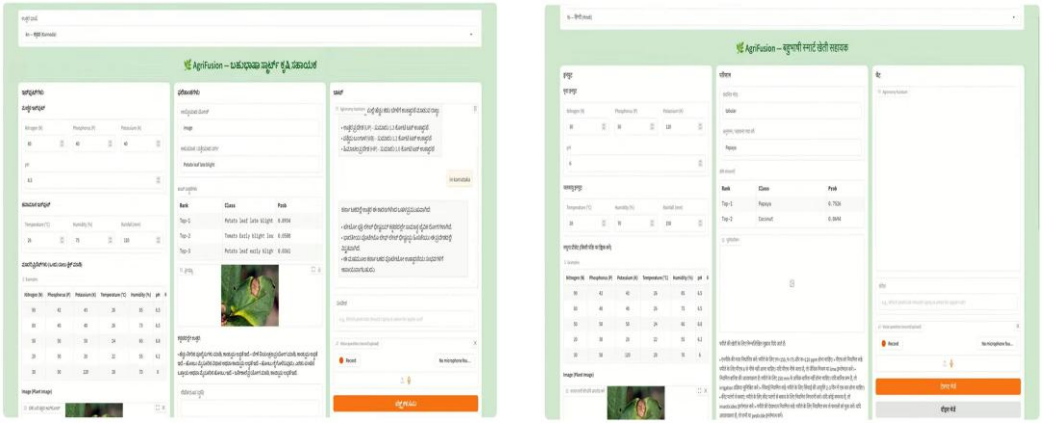
The Agrifusion application was deployed as an interactive web-based system integrating tabular crop prediction, image-based disease classification, multilingual reasoning, and speech interaction. Users provide soil–climate parameters or upload plant leaf images, and the system generates probabilistic predictions along with agronomic recommendations. The reasoning module produces contextual advisory outputs, including fertilizer guidance, irrigation scheduling, and disease management suggestions.

Figure 5 illustrates the Agrifusion multimodal application interface across different operational modes and languages. The platform supports six Indian languages— English, Hindi, Kannada, Telugu, Malayalam, and Tamil—through an integrated translation pipeline. A Whisper-based speech-to-text module enables voice input, while a text-to-speech engine generates audio advisory output, enabling fully multimodal agricultural interaction. A demonstration of the deployed system is available at [32].

The deployed application confirms the robustness of the agentic orchestration strategy. Tabular prediction, visual diagnosis, multilingual reasoning, and speech processing operate cohesively without requiring separate workflows. The system maintains consistent predictive outputs across languages while dynamically adapting the advisory narrative. The demonstration validates that the proposed framework is scalable, user-centric, and suitable for real-time agricultural advisory deployment.



(a) English interface showing tabular crop prediction and probability ranking (b) Kannada interface demonstrating image based classification



(c) Multilingual conversational reasoning in Kannada. (d) Hindi interface illustrating crop recommendation and advisory generation

Fig. 5: Agrifusion multimodal application deployment across languages and modalities. The system supports tabular crop prediction, image-based disease detection, multilingual conversational reasoning, and probability-based transparency within a unified interface.

5. Discussion

The experimental results demonstrate that the proposed agentic multimodal framework achieves strong predictive performance across both structured and visual agricultural tasks. Near-perfect macro F1-scores in crop recommendation indicate that nonlinear soil–climate interactions are effectively captured through optimized ensemble and deep architectures such as XGBoost [28] and transformer based models [30]. Similarly, disease classification performance on the PlantDoc dataset confirms that modern convolutional and transformer-based networks generalize robustly under real-field variability, including illumination shifts, occlusions, and background clutter.

While these quantitative results are competitive, the primary contribution of this work extends beyond accuracy alone. Many existing agricultural AI systems focus on isolated components of the decision pipeline. For instance, plant disease recognition frameworks emphasize classification performance and visual explainability, chatbot-based advisory systems concentrate on conversational interfaces [11] [16], and multilingual ASR research addresses speech adaptation for low-resource Indian languages [9]. However, these components are typically implemented as standalone modules without unified orchestration, cross-modal reasoning, or integrated explainability-aware advisory generation.

Recent multimodal large-model efforts such as Agri-LLaVA [23] demonstrate the potential of vision–language models for pest and disease understanding, while agent-oriented agricultural systems such as MA3[24] and [25] reflect a growing shift toward coordinated multimodal assistants. Nevertheless, most existing systems either emphasize vision–language

interaction or conversational retrieval without tightly coupling structured crop modeling, visual diagnosis, explainability propagation, and multilingual speech interaction within a single reasoning loop.

5.1 Implications

The implications of research work are as follows.

1. Agrifusion advances this direction through explicit agentic orchestration. First, it integrates structured tabular crop inference and image-based disease classification within a shared reasoning layer.
2. Second, interpretability artifacts—SHAP feature attributions [33] and Grad-CAM heat maps [34] are propagated directly into prompt construction, ensuring that advisory narratives are grounded in transparent model evidence.
3. Third, the framework completes the multimodal interaction cycle by incorporating speech-to-text and text-to-speech modules, enabling voice-first agricultural intelligence. This tight coupling preserves traceability from environmental inputs and lesion regions to probabilistic outputs
4. Finally, to actionable recommendations delivered in local languages.
5. Importantly, the deployment analysis confirms that the full multimodal loop tabular prediction, image classification, multilingual reasoning, and speech synthesis operates within practical latency constraints even on CPU-based hardware. GPU acceleration further reduces inference time for centralized or multi-user scenarios. These findings indicate that the proposed system is not merely a conceptual aggregation of AI components but a deployment-oriented prototype architecture adaptable to both decentralized rural environments and centralized advisory platforms.
6. Overall, the novelty of this work lies in the agentic coordination of heterogeneous AI modules into a unified, explainable, multilingual, and multimodal agricultural intelligence ecosystem. By binding predictive modeling, interpretability, reasoning, and bidirectional speech interaction within a single orchestrated pipeline, Agrifusion moves beyond modular AI tools toward a holistic decision-support framework suitable for real-world agricultural deployment.

5.2 Limitations

Despite many effectiveness of the proposed research work it has some of the limitations.

1. Dependency on Datasets and Generalization: The caliber and variety of training datasets have a significant impact on the performance of crop recommendation and disease diagnosis models. Applying the method to unseen geographic areas, crop kinds, or environmental variables may present difficulties.
2. Limited Validation in the Real World: Large-scale real-world deployment and long-term testing with farmers under various agricultural settings were few, despite the framework's impressive experimental findings.
3. Connectivity and Infrastructure Needs: Large language model reasoning and multilingual voice processing are examples of components that demand dependable internet access and compute power, which may limit their use in remote or low-resource locations.
4. Limitations on Explainability and Language: For low-resource languages and regional dialects, speech recognition and multilingual communication may be less successful.

6. Conclusion and future work

This work introduced **Agrifusion**, an agentic multimodal framework designed to deliver explainable, multilingual, and voice-enabled agricultural decision intelligence. Unlike conventional agricultural AI systems that operate as isolated predictive modules, the proposed architecture integrates structured crop recommendation, plant disease diagnosis, interpretability propagation, and large language model reasoning within a coordinated workflow. Experimental evaluation demonstrated near-perfect performance in tabular crop prediction, robust disease classification under real-world conditions, high-quality multilingual speech interaction, and practical end-to-end inference latency across deployment environments. Beyond predictive accuracy, the key contribution lies in the orchestration of heterogeneous AI components into a transparent and farmer-centric intelligence system. By embedding SHAP and Grad-CAM explanations directly into multilingual advisory generation, the framework preserves traceability from environmental inputs and lesion regions to actionable recommendations. The integration of speech recognition and synthesis further enables inclusive, voice-first accessibility for rural users. Overall, Agrifusion demonstrates the feasibility of scalable, explainable, and multimodal agricultural intelligence suitable for real-world deployment. The architecture establishes a foundation for next-generation AI-driven farming ecosystems that bridge predictive analytics, conversational reasoning, and practical field implementation.

Future Work

Future research will focus on extending the framework to incorporate real-time IoT sensor streams for dynamic soil and weather monitoring. Integration with satellite based vegetation indices and geospatial analytics could further enhance

region-specific crop advisory. Expanding the multilingual capabilities to additional low-resource Indian dialects and conducting structured user studies with farmers will provide deeper validation of usability and socio-economic impact. Additionally, lightweight model compression techniques may be explored to further optimize deployment on low-power edge devices.

Acknowledgements
The authors sincerely acknowledge the support of the Bioinformatics Institute A*STAR and School of Computer Science & Engineering, RV University, Bengaluru, India. The authors also thank the contributors of the publicly available datasets and open-source tools used in the development, experimentation, and evaluation of the Agrifusion framework.

Author's contribution statement
Arvind Channarayapatna Srinivasa and Revanasiddappa Bandi contributed equally to this work and share first authorship. Arvind C.S conceived the overall framework, designed the multimodal agentic architecture, implemented the experimental pipeline, conducted model development and evaluation, interpreted the results, and drafted the manuscript. Revanasiddappa Bandi contributed to system design, methodology development, dataset preparation, experimentation, validation, and manuscript refinement. Sriram Bindiganavile suresh contributed to conceptual review, multilingual validation, testing of the multi-agent workflows across different languages, result interpretation, and critical revision of the manuscript. All authors reviewed and approved the final manuscript.

Data Availability
The PlantDoc dataset used in this study is publicly available from its original published source [26]. The crop recommendation dataset used in this work is publicly available through Kaggle [27]. Processed data, trained model configurations, and additional experimental artifacts generated during the current study are available from the corresponding author upon reasonable request.

Funding
The authors received no specific funding for this work.

Ethics Approval
This study included a preliminary usability evaluation involving 28 voluntary participants who used the application for system assessment. Formal review by an Institutional Review Board or Ethics Committee was not obtained for this preliminary non-interventional usability study.

Consent to Participate
Informed consent was obtained from all participants prior to their inclusion in the study.

Conflicts of interest
There are no disclosed conflicts of interest for the authors.

References

[1] World Bank: Employment in Agriculture (% of Total Employment) – India. https: //data.worldbank.org/indicator/SL.AGR.EMPL.ZS?locations=IN. Accessed: 2026-02-17 (2026)
[2] Government of India, Ministry of Agriculture & Farmers Welfare: Agricultural Statistics at a Glance 2022. https://desagri.gov.in. Accessed: 2026-02-17 (2022)
[3] IPCC: Climate Change 2022: Impacts, Adaptation and Vulnerability. https://www. ipcc.ch/report/ar6/wg2/. Accessed: 2026-02-17 (2022)
[4] Tilman, D., Cassman, K.G., Matson, P.A., Naylor, R., Polasky, S.: Agricultural sustainability and intensive production practices. Nature 418, 671–677 (2002) https: //doi.org/10.1038/nature01014
[5] Savary, S., Willocquet, L., Pethybridge, S.J., Esker, P.D., McRoberts, N., Nelson, A.: The global burden of pathogens and pests on major food crops. Nature Ecology & Evolution **3**, 430–439 (2019) https://doi.org/10.1038/s41559-018-0793-y
[6] Wolfert, S., Ge, L., Verdouw, C., Bogaardt, M.-J.: Big data in smart farming: A review.

[7] Agricultural Systems **153**, 69–80 (2017) https://doi.org/10.1016/j.agsy.2017.01.023 Trendov, N.M., Varas, S., Zeng, M.: Digital Technologies in Agriculture and Rural Areas: Status Report. Accessed: 2026-02-17

[8] Jain, A., Bhowmick, S.: Asr models for low-resource indian languages: Comparative analysis and noise robustness. IEEE Access **13**, 12245–12256 (2025) https://doi. org/10.1109/ACCESS.2025.012245

[9] Singh, P., Sharma, R.: Asr adaptation for low-resource indian languages in agricultural contexts. Procedia Computer Science **230**, 344–351 (2023) https://doi.org/10.1016/ j.procs.2023.04.034

[10] Rehman, S., Thomas, A., Kumar, V.: Kisanqrs: Query response system for farmer call center analytics. Computers and Electronics in Agriculture 221, 108967 (2024) https://doi.org/10.1016/j.compag.2024.108967

[11] Sivakumar, T., Mohan, K.: A conversational chatbot framework for agricultural logistics management. International Journal of Agricultural and Environmental Information Systems 14(2), 45–59 (2023) https://doi.org/10.4018/IJAEIS.2023040104

[12] Suebsombut, J., Suranant, P., Wongsa, T.: Iot-enabled line chatbot for crop monitoring and advisory in smart farming. In: 2022 IEEE 7th International Conference on Smart Cloud (SmartCloud), pp. 111–117 (2022). https://doi.org/10.1109/ SmartCloud55606.2022.00023 . IEEE

[13] Reddy, V., Narayanan, P.: Agrohelper: A multilingual voice-based chatbot for farmer advisory. Applied Artificial Intelligence 38(5), 411–425 (2024) https://doi.org/10. 1080/08839514.2024.1011458

[14] Santhanavijayan, S., Manikandan, M.: Semantic enhancement of speech recognition for low-resource agricultural languages. Journal of Intelligent Systems 29(1), 167–179 (2020) https://doi.org/10.1515/jisys-2018-0497

[15] Pujar, A., Jain, D.: Gis-integrated conversational assistant for agricultural planning. In: Proceedings of the 2022 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 1024–1029 (2022). https://doi.org/10.1109/ IGARSS52255.2022.9884542 . IEEE

[16] Reddy, P., Balaji, D.: A text-based agriculture management chat assistant with nlp and rule-based reasoning. Journal of Computational and Cognitive Engineering **3**(1), 1–12 (2024) https://doi.org/10.1007/s43680-023-00205-1

[17] Arora, R., Mehta, D.: Agribot: Intelligent crop advisory using deep learning and nlp. Computers and Electronics in Agriculture 178, 105835 (2020) https://doi.org/10. 1016/j.compag.2020.105835

[18] Nikam, P., Deshmukh, S.: Agrodialogue: Llm-powered conversational ai for agricultural decision support. Expert Systems with Applications 241, 122854 (2024) https://doi.org/10.1016/j.eswa.2024.122854

[19] Bhuvaneswari, R., Kumar, S.: Ai-driven crop disease detection using cnn and transfer learning. Multimedia Tools and Applications 83(7), 18543–18561 (2024) https:// doi.org/10.1007/s11042-023-14701-3

[20] Gaikwad, S., Patil, A.: Agriqas: A question answering system for agricultural knowledge retrieval. In: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1472–1478 (2015). https://doi.org/ 10.1109/ICACCI.2015.7275812 . IEEE

[21] Suvarchala, S., Reddy, P.: Cottoncropqa: Domain-specific question answering system for cotton cultivation. Information Processing in Agriculture 10(3), 465–478 (2023) https://doi.org/10.1016/j.inpa.2022.12.003

[22] Shirisha, R., Naidu, P.: Multilingual virtual assistant for agricultural extension services. AI in Agriculture **13**, 100168 (2024) https://doi.org/10.1016/j.aiia.2024. 100168

[23] Wang, L., Jin, T., Yang, J., Leonardis, A., Wang, F., Zheng, F.: Agri-LLaVA: Knowledge-Infused Large Multimodal Assistant on Agricultural Pests and Diseases (2024). https://doi.org/10.48550/arXiv.2412.02158 . https://arxiv.org/abs/ 2412.02158

[24] Xu, Z., Xu, J., Zhang, M., Wang, P., Deng, C., Liu, C.-L.: Multimodal Agricultural Agent Architecture (MA3): A New Paradigm for Intelligent Agricultural Decision Making (2025). https://doi.org/10.48550/arXiv.2504.04789 . https://arxiv.org/abs/ 2504.04789

[25] Bhalerao, S., Rengarajan, A., Nadarajan, G., Shanmugam, A.S., Singh, P., Venugopal, S., Varma, K.: Krishi Sathi: A Multimodal and Intent-Aware Agricultural Assistant with Context Retrieval (2025). https://doi.org/10.48550/arXiv.2508.03719 . https: //arxiv.org/abs/2508.03719

[26] Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., Batra, N.: Plantdoc: A dataset for visual plant disease detection in the wild. In: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, pp. 249–253 (2020). https://doi.org/10.1145/3371158. 3371196

[27] https://www.kaggle.com/datasets/ samuelotiattakorah/agriculture-crop-yield, Accessed: 2026-02-17

[28] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794. ACM, ??? (2016). https://doi.org/10.1145/2939672. 2939785

[29] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)

https://doi.org/10.1162/neco.1997.9.8.1735
[30] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L ., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
[31] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. OpenAI Technical Report (2023)
[32] Google Cloud: Cloud Text-to-Speech Documentation. Available at: https://cloud.google.com/text-to-speech (2023)
[33] https://doi.org/10.1109/SSCI50451.2021.9659869, Accessed: 2026-02-17
[34] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (2017)
[35] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)