An Interpretable Multimodal Transfer Learning Decision-Support Framework for Multi-Crop, Multi-Disease Detection in Precision Agriculture

Deepali Shrikhande[1], Dr. Sushopti Gawade[2]

Abstract:
Despite the widespread use of image-only convolutional models for plant disease diagnosis to provide global food security, image variability at the field level, visually similar symptoms, and stress due to soil nutrients or moisture conditions, can cause loss of accuracy. This paper compares classical machine learning classification models to custom convolutional neural networks and pretrained transfer-learning models for multi-crop and multi-disease classification and also introduces a late-fusion multimodal decision support model that could integrate data about images, soil, and weather. Experiments were conducted on the PlantVillage dataset which had 54,305 RGB images belonging to 38 different crop-condition classes (43,456 training, 10,849 validation, and 10,849 test images) across 14 different crops. The accuracy, macro-precision, macro-recall and macro-F1 score were computed for the five classical classifiers (KNN, Random Forest, Extra Trees, SGD-linear SVM, and SVC-RBF using PCA-reduced features), three custom CNN variants, and four pretrained models (ResNet50, MobileNetV2, GoogleNet, and EfficientNetB7). Of the classical models, SVC with RBF kernel yielded the highest accuracy (83.58%) and MacroF1 (82.96%). In the case of the custom CNN models, the deeper they became and the more dropout the higher accuracy they gained, with CNN-V3 attaining 95.71% accuracy and 95.66% macro-F1. Transfer learning yielded the best results with MobileNetV2 (99.30% accuracy and macro-F1) outperforming ResNet50 (98.60% accuracy and macro-F1) and GoogleNet (97.83% accuracy and 97.46% macro-F1). The findings serve as the basis for the introduction of a multimodal system for soil forecasting integrating a MobileNetV2 image encoder, a residual MLP for soil features, and two dual LSTMs for 48-hour and 168-hour time-series of the weather. The interpretability, modularity and the sufficient resistance towards loss of sensor information for probability-level late fusion with 0.80, 0.10, 0.05 and 0.05 respectively make the framework appropriate for precision-agriculture advisory systems, until it is tested in the field.

[1]Research Scholar , Assistant Professor, Information Technology-INFT, Pillai institute of technology, Vidyalankar Institute of technology Mumbai, India, deepalishrikhande6@gmail.com
[2]Professor, Information Technology-INFT Vidyalankar Institute of technology Mumbai, India, sushoptiekrishimitra@gmail.com

Keywords: plant disease detection; multimodal transfer learning; decision support; precision agriculture; LSTM weather encoding; residual MLP.

1. Introduction
Plant diseases are a constant cause of yield and quality losses, especially in developing countries and are much more economical to prevent as soon as symptoms start to be noticed, rather than when they appear as a prominent visual problem. Manual scouting and laboratory verification, essential tools, are subject, time consuming, expensive, and not widely available in many production areas. In practice this is first noticed by growers/field workers when light, dust, cultivar specific morphology, water stress and nutrient imbalance are variable, and this causes an ambiguity which one can not identify from visual inspection.

Deep learning is transforming plant disease diagnostics by capturing disease classes directly from pixel information in a leaf image and frequently outperforming handcrafted-feature pipelines on curated benchmarks using convolutional neural networks (CNNs). There is, however, only a partial indication of plant health with an image. Pathogenic or nutritional causes, moisture stress, agrochemical injury, and disease promoting environmental conditions (leaf wetness, temperature, humidity, precipitation) may occur hours or days before symptoms are serious with a symptomatic leaf. A real-time implementation decision support system should be able to integrate leaf-level pattern recognition information, along with the agronomic information taken from soil and weather.

This paper reasons that multi-crop, multi-disease detection should be understood not merely as a conventional computer vision task, but as a multimodal inference problem. Image from leaves carries visible morphologic evidence; Soil attributes carry physiological and nutrient context; and Weather Time-series carries temporal disease pressure in terms of humidity, temperature, precipitation, and related microclimatic drivers. These complementary evidence streams can be fused in an interpretable and transparent way, producing a more informative and robust posterior to field variability and more actionable for operational decisions than an image-only posterior.

The distinction of this study from previous literature on image-centric plant diseases is that it makes the following contributions:

A systematic, head-to-head evaluation between 15 models along the three model families of classical machine learning algorithms, custom CNNs and pretrained transfer-learning backbones, on the 38-class, 14-crop PlantVillage corpus, with consistent preprocessing and macro-averaged evaluation protocol.

Development of a modular multimodal transfer-learning (MMTL) model architecture that brings together a MobileNetV2 image encoder and residual MLP for soil attributes and dual-window (48 h and 168 h) LSTM encoders for weather sequences.

A closed-form, interpretable equation with expert-prior weights, and an algorithmic specification (Algorithm 1) that is explicit enough to account for missing-modality renormalization for resilient field deployment.

A staged ablation and external-validation protocol, along with explainability (Grad-CAM), calibration, and responsible-deployment guidelines to inspire the empirical validation of the proposed framework on field-synchronized multimodal datasets.

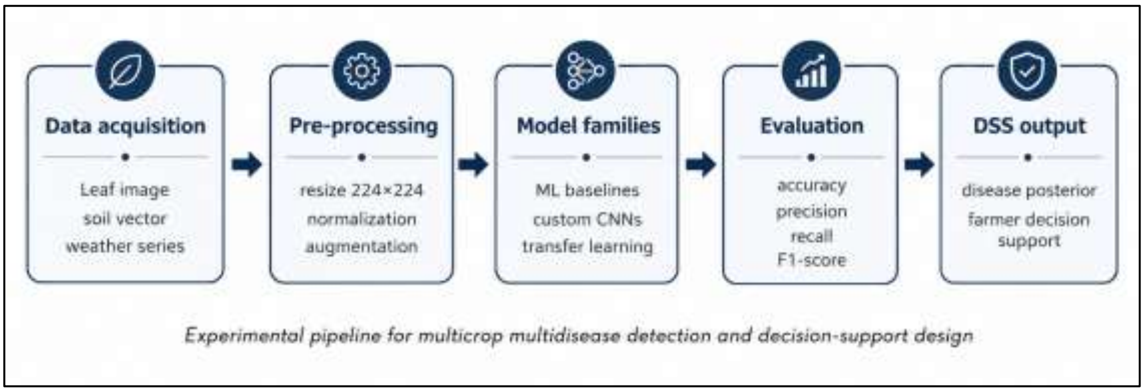

Figure 1. Experimental pipeline linking data acquisition, preprocessing, model evaluation, and decision-support output.

2. Literature Review
Those early computational solutions for plant disease detection were based on handcrafted features like color descriptors, texture measures and segmentation routines combined with traditional classifiers like SVM (support vector machines) and decision trees [1, 14, 17]. Such methods are still relevant as a form of reference usage since they are cheap to compute and, in certain fields, explainable and understandable. Their performance however is highly dependent on the quality of the pre-processing, consistency of illumination, and manually designed feature engineering and handcrafted descriptors do not necessarily extend to the diverse symptom morphologies found in a multi-crop environment.

A convolutional neural network can learn image representations on a hierarchy directly from pixels and thus minimizes the need for manual feature extraction. Accuracies obtained in various plant-disease studies using CNN surpasses those reported in other studies, especially when applied to well curated datasets like PlantVillage [2, 3, 9]. However, custom CNNs demand for appropriate network architecture, adequate pattern variety, regularization, and computational power, and shallow networks might underfit intricate disease patterns, while deep custom networks might be prone to overfitting when disease conditions are homogeneous or classes are imbalanced [7, 8].

There are a number of issues with training CNNs from scratch, which are tackled by transfer learning. The architectures trained on large-scale natural-image data can serve as reusable feature extractors capable of detecting natural image edges, textures, shapes, and higher-order visual patterns and are common choices for plant-disease diagnosis, such as ResNet50, MobileNetV2, GoogleNet/Inception, EfficientNet, and VGG variants (18, 19, 20). They perform differently in relation to the input resolution, the composition of the datasets, the augmentation strategy, the optimizer configured and deployment constraints. The depthwise-separable convolutions have low parameter count and computation, yet still possess good discriminative power, which makes MobileNetV2 especially appealing for in-the-field applications.

Notwithstanding these developments, there exists a common shortcoming in the literature: Most plant-disease models are image-based. Under controlled conditions a model can accurately classify a leaf lesion; it can be unsure when there is a variation in the field and when there is a similar class of symptoms that you can see. For example, early blight, leaf mold, and leaf spot of septoria are all characterized by the same color and texture cues in tomato [2, 6]. In such situations, an agronomic background may be helpful to clarify the situation as pathogens are limited by temperature, humidity, leaf wetness, rainfall, soil stressors, etc. [10, 21].

Agri-decision support systems (A-DSS) further extend the scope of classification to action, by linking predictions with recommendations. An alert is helpful only if it is timely, calibrated, interpretable and in line with field conditions [12, 22]. Multimodal systems are in line with the current development of digital agriculture in which sensor networks, satellite observations, mobile imagery, weather services and farm records are combined to take decision-based on data. [10, 22, 25] Multimodal learning, however, has well-established challenges such as those of sample synchronisation, missing streams, heterogeneity of feature scales, label consistency and interpretability [12, 23, 24]. Table 1 presents a summary of the main gaps that are considered as the motivating factors for the framework proposed.

Table 1 shows an overview of the key gaps that underlie the proposed framework.

Research gap	Technical consequence	Proposed response
Image-only diagnosis	Confuses visually similar diseases and may be overconfident under field conditions.	Fuse visual evidence with soil and weather context.
Context-only risk scoring	Raises false alarms when weather or soil is favourable but symptoms are absent.	Require image evidence while allowing context to calibrate probability.
Early-stage disease ambiguity	Borderline lesions may be missed because symptom expression is weak.	Use recent weather and soil stress to support borderline cases.
Missing data streams	End-to-end multimodal models may fail when a sensor stream is unavailable.	Set missing-modality weight to zero and renormalize remaining weights.
Single temporal window	Short events or long climatic patterns may be lost.	Use dual weather windows: 48 h and 168 h.

The scope and the emphasis of the present study are quite different: it is not intended to propose a new image classifier, but rather to offer (i) a controlled benchmark with multiple image families and a unified evaluation protocol and (ii) a formalization – in the sense of the architecture, fusion equation and deployment algorithm – of a multimodal extension that is explicitly introduced to be interpretable, modular and able to tolerate missing modalities (attributes) that are often discussed qualitatively in previous DSS literature but rarely specified at an implementation level of detail [12, 22, 23].

3. Materials and Methods

3.1 Dataset and Problem Formulation

The image-level experiments are conducted with the PlantVillage dataset consisting of 54,305 RGB images belonging to 14 crops, each containing 38 plant-health classes. There are four types of classes: Apple scab, Potato late blight, Tomato yellow leaf curl virus, or healthy leaf. The reported split consists of 43,456 training images, 10,849 validation images and 10,849 test images. This manuscript treats the validation and test partition the same size, as thus assigned, they are both considered to be streams of reported experimental evaluations rather than being parts of a single unique total.

The classification objective is designed to learn a mapping $f(X) \rightarrow y$, where X can be an image (image-only baselines) or a multimodal tuple $X = (X\_image, x\_soil, W\_full, W\_short)$ with the proposed framework, and $y \in \{1, \ldots, C\}$ with $C = 38$. In this case, the X_image is a Normalized RGB Leaf image (in $\mathbb{R}224\times224\times3$), the x_soil is a dense (11-dimensional) vector of soil attributes (such as pH, soil moisture, N, P, K, temperature, humidity) and the W_full is an hourly weather matrix, in $\mathbb{R}168\times11$, with 168 entries, one hour each and 11 columns, each representing a meteorological feature (e.g., wind, pressure, temperature), and the W_short is a 48-hour hourly weather matrix, 48 elements long, each element corresponding to an hour, there are 11 meteorological features.

The data consists of training, validation and testing data sets partitioned into fourteen crop groups: apple, blueberry, cherry, corn, grape, orange, peach, pepper bell, potato, raspberry, soybean, squash, strawberry, and tomato. To guarantee self-consistent model training and fair assessment of the model performance, an almost 80:10:10 data split was used. The largest group is tomato with 14,532 training images, 1,101 images each for validation and testing, and the soybean has 4,072 images each for training, validation, and testing. Multi-crop disease classification is provided by other crop groups such as grape, corn, apple, peach, pepper bell, potato, cherry, squash, strawberry and blueberry that also make significant contributions to the samples sent. Raspberry has the fewest numbers of training images (297) and there are 74 images for validation and testing. In general the data set exhibits a large degree of diversity in its crops, and their sampling, which allows the model robustness test, generalization ability test and classification performance test under imbalanced multi-crop.

3.2 Preprocessing and Experimental Protocol

Images were resized to 224×224 pixels to fit with the pretrained CNN backbones and custom CNN inputs, and pixel intensities were normalized to the range [0, 1]. The training-set was augmented with different geometric transformations, such as rotation, horizontal/vertical flipping, zooming, and minor spatial translation, to achieve robustness to the variations seen in the field for image acquisition.

Each model was trained as a multiclass classifier and optimized with categorical cross-entropy loss for CNN models and transfer-learning models, Adam was used for the optimization. Two metrics were used to monitor the training: validation loss and validation accuracy, and early stopping was used to combat overfitting and minimize over computation. The accuracy, precision, recall and F1-score metrics were used to evaluate performance, with the 38-class case highlighting the use of macro-averaging, which allows the most frequent classes to not unduly influence the evaluation and is thus more representative of the performance on the rare but agronomically significant disease categories.

Figure 2. Raw Input Images : After Resize and Before Normalization 224x224x3 RGB Tensors

3.3 Classical Machine-Learning Baselines

The first experimental family consists of five classical machine learning baselines, tested on processed image features. K-nearest neighbors (KNN) is based on Euclidean distance and five neighbors with distance-based weighting. Random Forest and Extra Trees are both with 100 trees. To keep computational cost low, the features are reduced to 1,000 components using PCA and the SVC is trained on those features, which is obtained by using a radial basis function (RBF) kernel. The linear SVM is implemented in SGD with hinge loss. These baselines measure value added by learned feature representations, and by transfer learning, over feature-engineered classical pipelines.

Table 3. Classical machine-learning models and their experimental role.

Model	Configuration	Purpose
KNN	k = 5, Euclidean distance, distance weighting	Simple non-parametric baseline; sensitive to feature noise.
Random Forest	100 decision trees	Robust ensemble baseline; moderate interpretability.
SVC (RBF)	PCA to 1,000 components; nonlinear RBF kernel	Strong nonlinear decision boundary; computationally intensive.
SGD	Linear SVM with hinge loss	Scalable and memory-efficient; sensitive to tuning.
Extra Trees	100 randomized decision trees, PCA-reduced features	Fast ensemble with high randomness and good generalization.

3.4 Custom CNN Architectures

Three versions of the custom CNN were assessed in order to understand the effect of the network capacity and regularization. CNN-V1 is composed of two convolutional blocks and a dense classifier. CNN-V2 introduces a third convolutional block to boost the representation capacity. CNN-V3 (using Keras functional API) consists of 4 convolutional blocks (same-padding), a dense layer and a high dropout rate (0.8). The aim was to see if deeper feature extraction with more robust regularisation can achieve better multicrop disease recognition in the course of the progression from V1 to V3.

The custom CNNs were trained from scratch to give a reference for dataset-specific learning capacity without the need of external pre-training. However, their representational power is limited by the quantity of data in the domain they are applied to, and they are expected to perform poorly on a high-class-count (38-class) problem relative to their pretrained counterparts, trained using transfer-learning models.But their representational power is limited by the amount of data in the domain to which they are applied and they are expected to perform poorly in a high-class-count (38-class) problem when compared to their pretrained counterparts derived from transfer-learning models.

Table 4. Summary of custom CNN architectures.

Variant	Architecture summary	Reported accuracy
CNN-V1	Conv2D(32)-MaxPool → Conv2D(64)-MaxPool → Flatten → Dense(256) → Softmax	84.80%
CNN-V2	Conv2D(32)-MaxPool → Conv2D(64)-MaxPool → Conv2D(128)-MaxPool → Dense(256) → Softmax	90.33%
CNN-V3	Conv2D(32/64/128/64, same padding)-MaxPool → Dense(256) → Dropout(0.8) → Softmax	95.71%

3.5 Transfer-Learning Models

Four pre-trained CNN architectures are adapted to the 40-class plant-disease problem and implemented transfer learning. ResNet50 features residual connections that enable deep representation learning; MobileNetV2 introduces inverted residual blocks and depthwise-separable convolutions for applying the network efficiently; GoogleNet (Inception) has inception modules to extract multi-scale features; and EfficientNetB7 has compound scaling to perform the network efficiently and consider both width, depth and input resolution. In both instances, the convolutional base is pretrained on visual information, and then a plant-disease classification head maps the output of the convolutional base to the plant-disease label space of 38 classes.

3.6 Proposed Multimodal Transfer-Learning (MMTL) framework

The proposed architecture is an extension of the best performing image-level model that emerged into a multimodal decision support architecture with four branches. This image branch features a pre-trained MobileNetV2 backbone, a lightweight dense head, and softmax output. A residual multilayer perceptron (MLP) with batch normalization and dropout that uses soil data. There are two LSTM-based encoders, the full-window encoder W_full with 168 hourly windows and the short-window encoder W_short with 48 hourly windows, called the weather subsystem. A C-dimensional class-posterior vector is produced for each branch, which corresponds to the 38 disease categories.

The fusion is carried out on the probability level with a fixed convex combination. Assume that the softmax-normalized class-posterior vectors from the image, soil, full-weather and short-weather branches are denoted as p_image, p_soil, p_weather-full and p_weather-short, respectively. The posterior p is fused as follows:

$$p = 0.80 \cdot \text{pimage} + 0.10 \cdot \text{psoil} + 0.05 \cdot \text{pweather-full} + 0.05 \cdot \text{pweather-short}$$

with the last class being the one with the maximum argument from p(c), i.e. $\hat{y} = \mathrm{argmax}_c\,(p(c))$ The sum of the four weights is one, making p an unnormalized probability vector over the 38 classes, with each branch's posterior being individually softmax-normalized.

There are three reasons why late fusion is used.There are three reasons for taking such an approach with late fusion. The first is that it is interpretable: every modality provides an explicit inspectable probability component to the final decision. Second, it is modular: any single branch can be trained, replaced or ablated separately without training the rest. Third, it is tolerant of missing data: when it is not possible to access a stream on inference time, the weights of that stream may be set to zero and the remaining weights re-normalized (Algorithm 1, step 6) which is critical when deploying in field settings where sensor and/or weather feeds may be intermittent.

Table 5. Modality-specific encoders and probability-level fusion weights.

Modality	Encoder	Information captured	Fusion weight
Image	MobileNetV2 + dense classifier	Leaf symptom morphology and texture	0.80
Soil	Residual MLP with batch normalization and dropout	pH, moisture, N, P, K, temperature, humidity, and related attributes	0.10
Full weather	Stacked LSTM, 168 h × 11 features	Longer-horizon disease-pressure and climate context	0.05
Short weather	Compact LSTM, 48 h × 11 features	Recent humidity, rainfall, and temperature dynamics	0.05

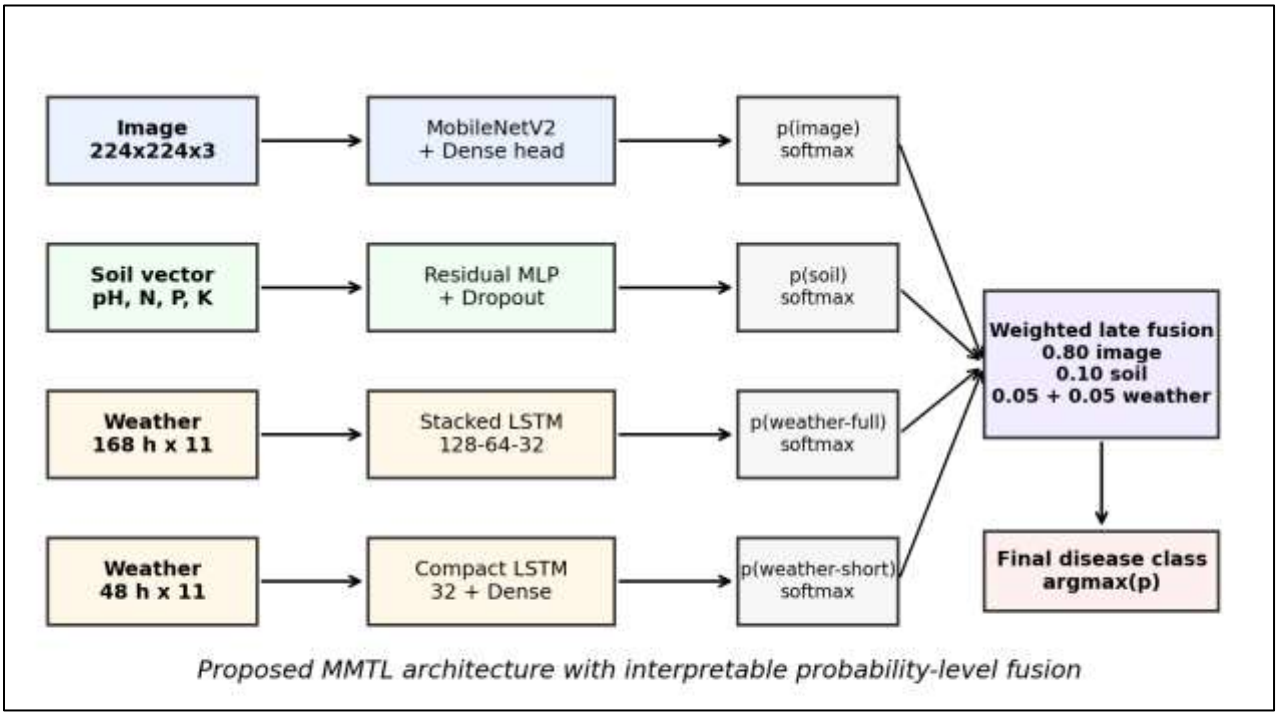

Figure 3. Proposed multimodal transfer learning architecture with image, soil, and dual weather encoders.

3.7 Multimodal Data Alignment

Sample alignment across heterogeneous information sources is one of the key engg. needs for multimodal learning. For the proposed FusionSequence generator design, one sample-index array is kept across all modalities, which is considered as a sample-index array for a canonical design. The image labels provide the reference source, and the soil vectors and weather matrices are referred to the same sample index. Weather sequences are pregenerated and stored in dense, indexable sequences to be accessed deterministically during training, avoiding hidden mismatches between the sequences of independent generators or of streaming iterators since they are not consumed at the same speed.

The generator shuffles with epoch granularity without losing any correspondence across modalities and reinitializes an underlying iterator safely as soon as an iterator's end is reached. The soil and weather features are encoded as a float32 tensor and the labels are transformed into one hot categorical encoding. Any lack of length match or missing record should be diagnosed from the log and is considered critical; otherwise, misaligned performance during training can yield apparent failures or poor performance stability that can be difficult to post mortem.

3.8 Evaluation Metrics and Mathematical Definitions

Overall accuracy is inadequate for a 38 class diagnosis problem as classes are not equally distributed. Precision, recall and F1-score give complementary information of the false-positive and false negative behaviour. In a class c, precision is the ratio of the number of correct predictions to the total number of predictions made in that class c, and recall is the ratio of the number of correctly recovered class-c samples to the total number of class-c samples. The F1-score is the harmonic mean of the recall and precision and especially useful in situations where relative costs of false positives and false negatives need to be taken into account.

The following are basic metric terms:

$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$

$\text{Precision} = TP / (TP + FP)$

$\text{Recall} = TP / (TP + FN)$

$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$

In the multiclass setting, these quantities are calculated class-wise, and then macro-averaged. Because rare but aggravating disease cannot be hidden by excellent performance on overabundant, healthy or common-crop disease categories, macro-averaging is recommended for disease surveillance.

In the case of the proposed multimodal system, the fused posterior p is computed against the one-hot target vector y and measured by categorical cross-entropy:

$CE(y, p) = - \Sigma c\ y(c) \log p(c)$

If p is generated from a convex combination of calibrated branch posteriors, the fused probability is interpretable. Further evaluation of the fused posterior needs to be done in terms of top-k accuracy, expected calibration error (ECE), Brier score, and confidence-threshold analysis in operational environments. These measures define the timing of the system's recommendations and suggestions for expert confirmation, and when it asks for a recaptured image.

Table 6. Evaluation metrics used for multiclass plant disease detection.

Metric	Definition	Interpretation for plant disease DSS
Accuracy	(TP + TN)/(TP + TN + FP + FN)	General correctness across all classes; sensitive to class imbalance.
Precision	TP/(TP + FP)	Measures reliability of positive predictions for a disease class.
Recall	TP/(TP + FN)	Measures ability to recover true disease cases; important for early warning.
F1-score	2PR/(P + R)	Balances precision and recall; preferred when both error types matter.
Cross-entropy	minus sum over c: y(c) log p(c)	Training loss for multiclass posterior estimation.

3.9 Algorithmic Specification of Late-Fusion Inference

The proposed inference procedure is summarized in Algorithm 1 from an implementation point of view. The algorithm is intended to be used modularly, that is, each modality specific branch is capable of running on its own and the fusion part sums the available posteriors once the streams are available and they are found to be legitimate.

Table 7. Algorithm Late-fusion inference procedure for the proposed multimodal decision-support system.

Step	Operation
1	Acquire X(image), x(soil), W(full), and W(short) for the same crop sample and timestamp.
2	Resize X(image) to 224x224x3 and normalize pixels; validate soil ranges and weather timestamps.
3	Compute p(image) = MobileNetV2_head(X(image)).
4	Compute p(soil) = ResidualMLP(x(soil)) if soil data are available; otherwise mark soil stream missing.
5	Compute p(weather-full) = LSTM_full(W(full)) and p(weather-short) = LSTM_short(W(short)) if weather windows are available.
6	Set missing-stream weights to zero and renormalize remaining weights so that sum of w(m) = 1.
7	Compute p = sum over m of w(m)p(m) and return top-1 class, top-k alternatives, confidence, and modality contributions.
8	If max(p) is below a deployment threshold, request recapture, expert validation, or laboratory confirmation.

4. Results

All results are reported on the test partition not used for training (10,849 images) and are seen with accuracy, macro-precision, macro-recall and macro-F1 (as defined in Section 3.8). early stopping is used as a validation loss for all models, with the reported test metrics extracted from the model at the point of minimum validation loss. If more than one training run is carried out in this manuscript, mean values should be shown, and standard deviations at least over three random seeds as well as pairwise significance tests (e.g., McNemars' test or a paired bootstrap based on correctness per sample) should be reported and can be used to support the claim of superiority between closely performing models like MobileNetV2 and ResNet50.

4.1 Classical Machine-Learning Performance

The performance hierarchy of classical machine-learning baselines is in line with expectations. The raw features and shallow processed features give poor KNN classification performance (accuracy 50.27%, macro-F1 46.37%) in the task of 38-way crop-disease discrimination. The top three models—Random Forest, SGD, and Extra Trees—tend to be pretty consistent at roughly 63–64% accuracy, and that's actually a statement of how limited these models are of their own, when they are not learning end-to-end representations. The performance provided by SVC with RBF kernel is significantly better than the other classical models with an accuracy of 83.58% and macro F1 of 82.96%, which shows the importance of nonlinear decision boundaries after dimensionality reduction via PCA.

Table 8. Performance metrics for classical machine-learning models.

Model	Accuracy	Precision	Recall	F1-score
KNN	0.5027	0.5850	0.5027	0.4637
Random Forest	0.6327	0.6339	0.6327	0.5955
SVC (RBF)	0.8358	0.8315	0.8358	0.8296
SGD	0.6369	0.6210	0.6369	0.6227
Extra Trees	0.6405	0.6640	0.6405	0.6049

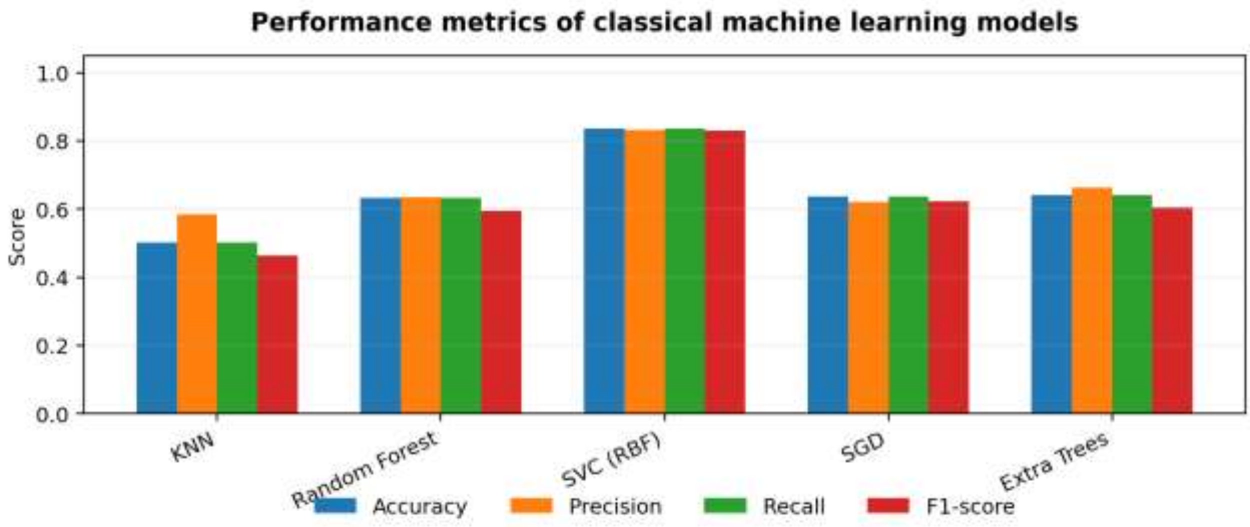

Figure 4. Classical machine-learning performance across accuracy, precision, recall, and F1-score.

4.2 Custom CNN Performance

As the depth of the CNN and the extent of regularization increase, it consistently enhances the CNN's performance. CNN-V1 achieves an accuracy of 84.80% already close to the SVC baseline, despite being very simple, having only 2 blocks, and being trained from scratch. CNN-V2 performs better at 90.33%, suggesting deeper convolutional depth helps to detect more discriminative symptom patterns. CNN-V3 gives the best results in terms of custom-CNN performance: 95.71% accuracy and 95.66% macro-F1, despite the high dropout rate (0.8) which seems to result in better generalization and fewer signs of overfitting, but also longer training convergence.

Table 9. Performance metrics for custom CNN variants.

Model	Accuracy	Precision	Recall	F1-score
CNN-V1	0.8480	0.8578	0.8480	0.8479
CNN-V2	0.9033	0.9042	0.9033	0.9015
CNN-V3	0.9571	0.9580	0.9571	0.9566

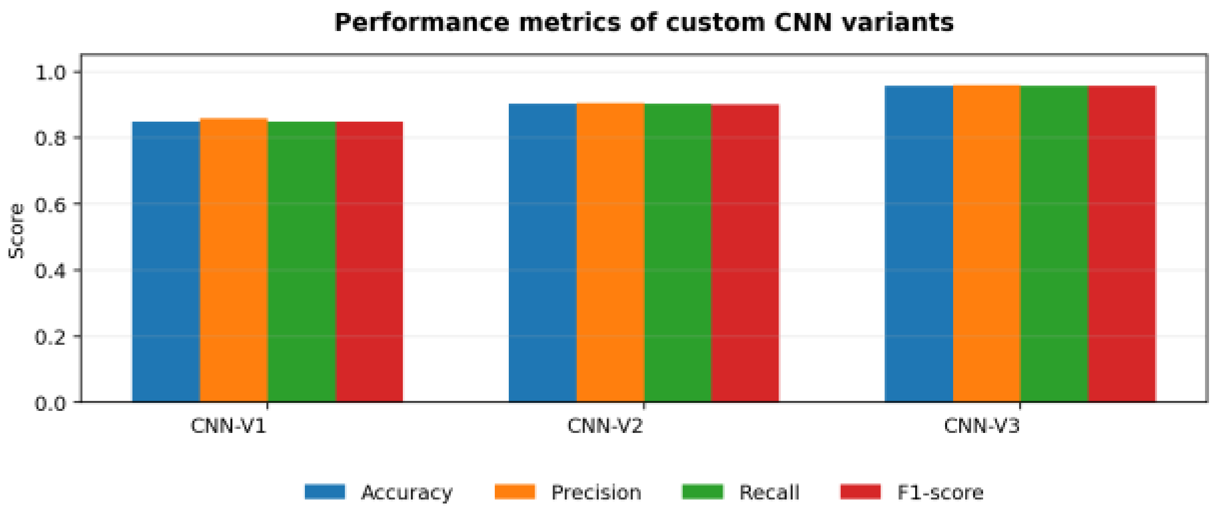

Figure 5. Custom CNN performance improves from shallow sequential design to regularized functional CNN.

4.3 Transfer-Learning Performance

Transfer learning gives the best results at the image level among all families assessed. MobileNetV2 achieves 99.30% accuracy, 99.32% precision, 99.30% recall, and 99.30% macro-F1. ResNet50 comes in very close behind with an accuracy of 98.60%, recall of 98.60% and an F1-score of 98.60%. The GoogleNet model has a high accuracy of 97.83% and macro-F1 of 97.46%, which is still a competitive result. While EfficientNetB7 performs poorly at 33.90%, probably due to its default architecture, it highlights that different experiments may require different configurations of EfficientNetB7. This may be due to inadequate fine tuning of the upper layers, a mismatch between the resolution or preprocessing of the input and how EfficientNet is designed to work, instability when training due to a larger number of parameters in the network, or the need to optimize certain hyperparameters for the specific architecture (like a larger learning-rate warmup, a smaller batch size etc.).

Table 10. Performance metrics for pretrained transfer-learning models.

Model	Accuracy	Precision	Recall	F1-score
ResNet50	0.9860	0.9863	0.9860	0.9860
MobileNetV2	0.9930	0.9932	0.9930	0.9930
EfficientNetB7	0.3390	0.3940	0.3390	0.3340
GoogleNet	0.9783	0.9723	0.9783	0.9746

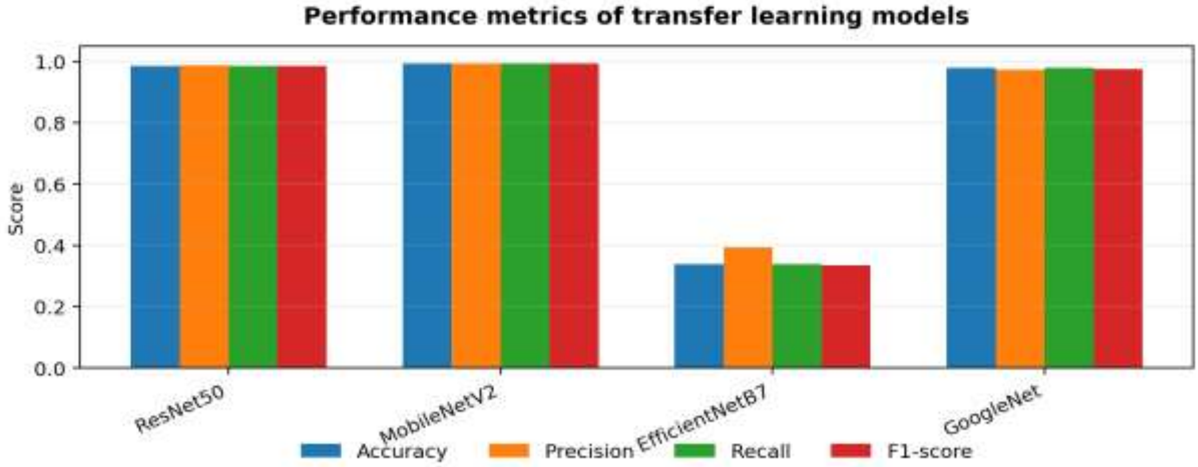

Figure 6. Transfer learning model comparison, showing MobileNetV2 as the best image-level model.

4.4 Overall comparison and confusion-pattern interpretation

In precision-agriculture applications, which are frequently constrained by mobile or edge hardware, Reporting of plant-disease classifiers is now becoming more important than accuracy, as it is crucial to reduce the cost of such a system. The parameter counts and relative inference characteristics listed are indicative based on published architectural specifications, and are subject to change pending future revisions, with the measured inference latency and memory footprint on the target deployment hardware (e.g., representative smartphone, edge gateway) to be reported.

As a whole, there is evident progression from classical learning, to custom deep learning and finally to transfer learning. The best classical model, SVC, achieves a macro-F1 of 82.96%, CNN-V3 achieves 95.66% and MobileNetV2 achieves 99.30%. It reveals that convolutional representations learned on the source domain are crucial for fine-grained disease classification across a 38-class, 14-crop problem and that pre-trained architectures enable the transfer of highly transferable image features between a source and target domain that is quite different from each other (natural images vs. leaf images).

Based on the qualitative assessment of the error matrix, it can be concluded that the errors of the model are grouped together for visually similar diseases, including tomato and tomato-related classes, including early blight, late blight, leaf mold. These confusions are biologically plausible because there are several fungal and bacterial diseases that induce similar lesion geometry, color gradients, necrotic margins and chlorotic halos. In these situations, the proposed soil and weather branches are more than just supplemental evidence; they also give some indication of the relative merits of the field's two potential diagnoses, encouraging the multimodal extension.

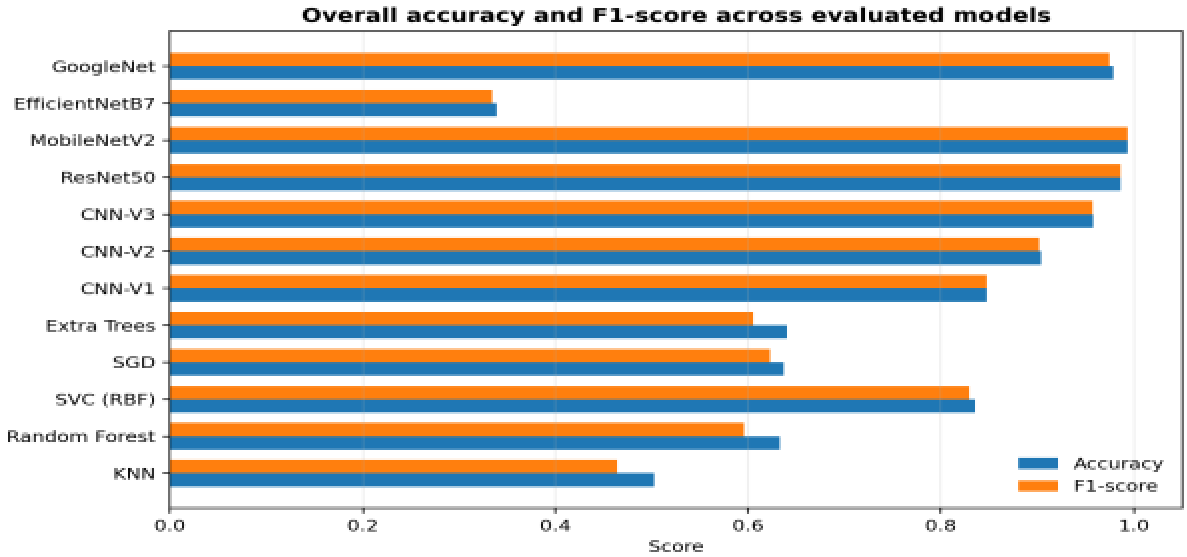

Figure 7. Overall accuracy and F1-score across all evaluated model families.

5. Discussion

The results confirm three technical observations. First classical methods can give meaningful baselines but are limited by the quality of the feature extraction; the good performance of the classical SVC suggests that nonlinear decision boundaries are better, but the computational cost and lack of the end-to-end feature learning make the results less scalable to larger or changing class sets. Second, the results from the custom CNNs demonstrate that features obtained by hierarchical convolution better suit the disease recognition than engineered features, and that the difference between CNN-V1 and CNN-V3 indicates that the depth of the architecture and the use of regularization are crucial in a multicrop recognition task with 38 classes. Third, pretrained transfer learning is the best image-level solution, especially when high accuracy is needed for a data set that is too small or lacks the necessary computational resources to train large CNNs from scratch.

MobileNetV2 is recognized as the most suitable backbone for the proposed decision-support framework for its ability to achieve high accuracy at the image level (99.30%) and relatively efficient deployment (Table 10). For precision agriculture applications, models often must execute on a mobile device, a low-cost edge gateway or intermittently connected rural location. The lower latency, energy usage, and network connectivity requirements of lightweight networks make them suitable for cloud infrastructure, whereas a heavier model like EfficientNetB7 or ResNet50 could be a better choice. The performance results reported, then, justify MobileNetV2 as a viable and technically viable option for the image modality of the envisaged multimodal system.

The not-so-stellar performance of EfficientNetB7 (33.90% accuracy) is not simply ignored as a curiosity, but seen as a significant scientific discovery. EfficientNet variants are typically quite competitive, but they are highly dependent on input resolution, pre-processing, learning-rate schedule, batch-size and tuning the depth. If trained with settings more appropriate to smaller networks, a compound-scaled architecture can be unstable and/or sample-inefficient. This finding also highlights the need for hyper-parameter optimization and ablation studies prior to determining the unsuitability of a given architecture for plant-disease detection and recommends future work to report a dedicated study on fine-tuning a hyper-parameter set, specifically progressive unfreezing, resolution matching to 600x600 as done for EfficientNetB7, and reduced learning rate, to isolate architecture effects from hyper-parameter effects.

The multimodal fusion design proposed directly bridges the gap between controlled image classification and field-grade decision support. An image only model can learn from an image, but not determine whether humidity, rainfall, soil moisture, or pH at the time the image was taken increase or decrease the likelihood of a disease. In the late fusion framework, the evidence of visualization will be more influential when symptoms are clear (weight 0.80) and the evidence of soil/weather will be more influential when the symptoms are early or ambiguous. This can be especially beneficial for similar diseases, when nutrient stress is confounding, or when there is a seasonal shift in disease occurrence.

Probability-level fusion is further advantageous in that it is interpretable. Since every modality will have a different class distribution, the system can give the details about the weightage of the image, soil and weather evidences in the final prediction. This transparency is operationally significant for a decision support system that is developed for farmers: a message saying for example that the humidity has been high for the last 48 hours and the symptoms of the leaves correspond to the symptoms of a certain disease is much more useful to the farmer than an unexplainable class label.

The framework also provides for graceful degradation in case of partial data availability. In the absence of soil sensors, the soil weight can be considered as zero and the system can be temporarily run as an image-plus-soil model; in the absence of weather-service access, the system can be temporarily run as an image model; and in the absence of any weather-service access, the system can fall back to the best unimodal model (MobileNetV2). This is better than early-fusion architectures where raw modality inputs are combined prior to a common encoder, because these architectures need all inputs to be in place during inference and need retraining or imputation schemes if a modality is missing.

5.2 Infrastructure for a Farmer-Oriented DSS

The deployable DSS can be broken down into four layers. Leaf images are captured with a mobile application, soil measurements are manually entered or obtained from sensors, and weather sequences are received from local stations or weather APIs. The preprocessing layer resize and normalize images, validates ranges of soil attributes, and merges together the weather observation into 48-hour and 168-hour matrix. The inference layer is where the four modular encoders are activated and where weighted late fusion (as per Algorithm 1) is done. Lastly, the fused posterior is translated to the final disease probability, a confidence category, recommended verifications steps, and agronomic alerts.

In low-resource environments, MobileNetV2 branch can be deployed on-device, and soil and weather processing could be done on-device or in the cloud based on the connectivity. Update the model can be done periodically as new field images and expert labels become available. All predictions should be recorded in a production system with metadata such as crop, location, date, environmental values, model version, confidence, and user feedback, and these records allow for continuous learning, model auditing, and recalibration for various regions over subsequent growing seasons.

5.2.1 Limitations of the study

The major drawback is that the images in PlantVillage are cleaner and more controlled than typical field images. Results from the real-world deployment therefore need to be assessed using field-captured images containing background clutter, multiple leaves per image, occlusion, varying illumination, and a mix of symptoms. Secondly, the classes are imbalanced, for example, the class of Tomato yellow leaf curl virus has much more samples than the 38 healthy or rare-disease classes (Table 2); in future reports, the macro-averaged metrics, per-class recall and confusion matrices should be emphasised and reported for all 38 classes.

Another constraint is that the multimodal model (as defined in Section 3.6) necessitates synchronised soil and weather data. The architecture, fusion equation and dataflow are formalized in the present manuscript but the final multimodal performance is to be validated through a field dataset for which all modalities are gathered for the same crop, site, and time. If there isn't a true "temporal" and "spatial" match, the soil or weather data can be a source of noise rather than informative context. The fourth limitation is that the fixed fusion weights (0.80/0.10/0.05/0.05) represent an expert prior. While these weights are interpretable, they might not be best across crops, seasons, or disease type; learnable gating, learnable attention-based fusion, Bayesian calibration, and uncertainty estimation are promising extensions that can be explored in the ablation matrix below (Table 11).

The fifth limitation of single-dataset benchmark studies is that there is no evidence of generalization across datasets. The results of transfer learning reported (Table 9) are from PlantVillage; transfer learning performance on field data sets other than PlantVillage (such as PlantDoc or region-specific field corpora) would be expected to be less and should be reported in future work to bound the external validity of the accuracy figures reported here.

5.3 Ablation and validation strategy

There should be a staged ablation study for empirical validation of the proposed MMTL framework. The first stage is based on the image branch alone, achieving the highest unimodal reference (MobileNetV2, 99.30% accuracy). The second stage incorporates soil attributes, to determine if the nutrient and moisture context has a beneficial effect on difficult classes. The third step is to add short window weather only, full window weather only and a combination of both. The fourth stage evaluates the relative performance of fixed fusion, learnable gating, attention-based fusion and stacking of calibrated probabilities, to see if the expert-prior weights in Section 3.6 are competitive with learnable alternatives.

The aggregation level of an ablation is not enough for a rigorous ablation. It should return per-class recall figures in the case of economically important diseases, confusion matrices for look-alike diseases, calibration curves, failure cases plotted vs the confidence threshold, and representative failure cases. For instance, if the weather stream recall for tomato late blight is improved with the 48-hour weather stream, then the benefit is cross checked with the history of humidity and rainfall to verify agronomic plausibility. Likewise, if soil characteristics provide extra tests and additional false positive look-alikes, the model explanation should point to the soil parameters or indicators (such as nitrogen deficiency indicators) that lead to the change.

External validation is crucial as field performance can be overestimated by benchmark data. A recommended validation design is a collection of co-timed samples from different cultivars, devices, seasons and farms. A minimum of one image, a time stamp, crop identity, a disease label that has been verified by experts, soil measurements within a specified time window, and a weather history in the hour.At least one image, a time stamp, crop identity, an expert verified disease label, soil measurements within a defined time window, and a weather history in the hour. The model should then be tested on farms and seasons not encountered in training to quantify the domain shift and generalization and results should be presented in terms of the metrics in Table 6 broken down by farm and season.

Table 11. Recommended ablation matrix for validating the proposed multimodal system.

Ablation setting	Fusion form	Purpose
Image only	p(image)	Strongest visual reference; quantifies benchmark image performance.
Image + soil	0.90 p(image) + 0.10 p(soil)	Tests whether soil stress improves visual ambiguity resolution.
Image + 48 h weather	0.90 p(image) + 0.10 p(weather-short)	Tests immediate disease-pressure contribution.
Image + 168 h weather	0.90 p(image) + 0.10 p(weather-full)	Tests longer climatic context.
All modalities	0.80 p(image) + 0.10 p(soil) + 0.05 p(weather-full) + 0.05 p(weather-short)	Full proposed decision-support model.
Learnable fusion	g_m(x) p_m	Tests whether adaptive weights outperform fixed expert priors.

5.4 Explainability, Calibration and Responsible Recommendations

Explainable at two levels is a precondition for a farmer facing diagnostic system. The mapping methods of class activation, like Grad-CAM, can visualize the parts of the leaf where the model made its prediction, which may reveal overreliance on backgrounds, leaf edges, shadows, or artifacts from a specific dataset. The system should make available the probabilities at the branch level and their relative weights, so that the end user can be made aware of the predicted disease as well as if it was mostly based on the disease weather or soil stress.

Proper calibration is essential for good decision support. Even a model that is 99% accurate can yield poorly calibrated confidence scores if applied to a new region or a new device. The system should thus include temperature scaling or a different method of calibration based on an independent validation set. Confidence thresholds can then be translated to action categories: high confidence then an immediate advisory, medium confidence then monitoring and image recapture, and low confidence then expert review. This staged reaction minimizes over-chemicalization or under-chemicalization of the crop.

Ethically, the DSS should not be seen as an independent agricultural decision making centre, but rather a support mechanism. It should express doubt, allow user feedback for constant testing, and not give too many recommendations. If it is a high impact disease or a high cost intervention (e.g., broad-spectrum fungicide spraying), confirmation should be suggested by using an agronomist inspection or laboratory testing. Data protection measures for farm locations, image metadata and production data also need to be applied in a responsible way, in line with regional data protection laws.

6. Conclusion

This paper provided a technical, journal-style evaluation of three model families, classical machine learning, custom CNNs and pretrained transfer learning, for multi-crop, multi-disease detection on the PlantVillage benchmark, and formalized a multimodal transfer-learning (MMTL) framework that extends the best performing image-level model into a field-deployable decision-support system. The evaluation reveals that the classical machine-learning baselines are useful for setting a performance floor but not enough to tackle fine-grained recognition of plant diseases in 38 classes. The performance gain is achieved by significantly deepening the network structure and using regularization, and the best image-level performance is obtained with pretrained transfer learning, with the best macro-F1 of 99.30% and accuracy of 99.30% using MobileNetV2. The results support the adoption of a light weight pretrained visual backbone as the backbone of a farmer friendly diagnostic system.

The proposed MMTL framework introduce an interpretable, probability level late-fusion equation, which combines the weights from the expert prior, image classification and soil attributes to create an agronomically grounded decision support system. The resulting design is (i) modular, (ii) explainable, (iii) generates a design that is robust against missing data streams through weight renormalization, and (iv) designed to be applicable in the context of precision-agriculture decision support in farmer friendly way, as will be shown by the field-synchronized empirical testing in Sections 5.2 and 5.3.

Future work should focus on four aspects: (i) validation of the complete fusion model on multi-farm, multi-season field datasets with synchronized image, soil, and weather streams; (ii) assessing the advantages and disadvantages of the fixed-prior fusion approach compared to learnable gating, learnable attention-based fusion and learnable stacked calibrated probabilities; (iii) quantification of the uncertainty of the predictions and learnable calibration of probabilities across locations, seasons, and devices; and (iv) integration of explainable-AI techniques, e.g., Grad-CAM for visual symptom localization and reporting scores per-modality of contribution into the deployed advisory interface. These extensions open the door for practical, evidence-based and reliable, sustainable crop-disease management since the multimodal transfer learning can be a basis.

Data Availability Statement

The data used in this study (PlantVillage) are publicly available at https://plantvillage.psu.edu/ [15]. The collection of soil and weather data was not carried out within the scope of this study, but plans for collecting and sharing of soil and weather data for field synchronised multimodal data will be described in a follow-up study.

References

[1] Wani, J. A., Sharma, S., Muzamil, M., et al. Machine learning and deep learning based computational techniques in automatic agricultural diseases detection: methodologies, applications, and challenges. Archives of Computational Methods in Engineering, 29, 641-677, 2022.

[2] Sakkarvarthi, G., Sathianesan, G. W., Murugan, V. S., Reddy, A. J., Jayagopal, P., and Elsisi, M. Detection and classification of tomato crop disease using convolutional neural network. Electronics, 11, 3618, 2022.

[3] Geetharamani, G., and Arun Pandian, J. Identification of plant leaf diseases using a nine-layer deep convolutional neural network. Computers and Electrical Engineering, 76, 323-338, 2019.

[4] Selvaraj, M. G., Vergara, A., Ruiz, H., et al. AI-powered banana diseases and pest detection. Plant Methods, 15, 92, 2019.

[5] Ahmed, A. A., and Reddy, G. H. A mobile-based system for detecting plant leaf diseases using deep learning. AgriEngineering, 3, 478-493, 2021.

[6] Pallagani, V., Khandelwal, V., Chandra, B., Udutalapally, V., Das, D., and Mohanty, S. P. dCrop: A deep-learning based framework for accurate prediction of diseases of crops in smart agriculture. IEEE iSES, 2019.

[7] Kabir, M. M., Ohi, A. Q., and Mridha, M. F. A multi-plant disease diagnosis method using convolutional neural network. In Computer Vision and Machine Learning in Agriculture, Springer, 2021.

[8] Roy, A. M., and Bhaduri, J. A deep learning enabled multi-class plant disease detection model based on computer vision. AI, 2(3), 413-428, 2021.

[9] Ferentinos, K. P. Deep learning models for plant disease detection and diagnosis. Computers and Electronics in Agriculture, 145, 311-318, 2018.

[10] Antle, J. M., Jones, J. W., and Rosenzweig, C. Next generation agricultural system models and knowledge products: synthesis and strategy. Agricultural Systems, 155, 179-185, 2017.

[11] Lee, S. H., Goëau, H., Bonnet, P., and Joly, A. New perspectives on plant disease characterization based on deep learning. Computers and Electronics in Agriculture, 170, 105220, 2020.

[12] Zhai, Z., Martinez, J. F., Beltran, V., and Martinez, N. L. Decision support systems for agriculture 4.0: survey and challenges. Computers and Electronics in Agriculture, 170, 105256, 2020.
[13] Ingram, J., Maye, D., and Bailye, C. What are the priority research questions for digital agriculture? Land Use Policy, 114, 105962, 2022.
[14] Demilie, W. B. Plant disease detection and classification techniques: a comparative study of the performances. Journal of Big Data, 2024.
[15] PlantVillage. PlantVillage plant disease image dataset, Pennsylvania State University. Available online: https://plantvillage.psu.edu/.
[16] Aldhyani, T. H., Alkahtani, H., Eunice, R. J., and Hemanth, D. J. Leaf pathology detection in potato and pepper bell plant using convolutional neural networks. ICCES, 2022.
[17] Panigrahi, K. P., Das, H., Sahoo, A. K., and Moharana, S. C. Maize leaf disease detection and classification using machine learning algorithms. Progress in Computing, Analytics and Networking, Springer, 2020.
[18] Prodeep, A. R., Hoque, A. M., Kabir, M. M., Rahman, M. S., and Mridha, M. F. Plant disease identification from leaf images using deep CNN's EfficientNet. DASA, 2022.
[19] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. A survey on deep transfer learning. International Conference on Artificial Neural Networks, 2018.
[20] Too, E. C., Yujian, L., Njuki, S., and Yingchun, L. A comparative study of fine-tuning deep learning models for plant disease identification. Computers and Electronics in Agriculture, 161, 272-279, 2019.
[21] Chlingaryan, A., Sukkarieh, S., and Whelan, B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. Computers and Electronics in Agriculture, 151, 61-69, 2018.
[22] Fountas, S., Espejo-García, B., Kasimati, A., Mylonas, N., and Darra, N. The future of digital agriculture: technologies and opportunities. IT Professional, 22(1), 24-28, 2020.
[23] Shrivastava, S., Bansal, P., Gupta, V., and Kumar, A. Effective multi-crop disease detection using pruned deep learning models. Expert Systems with Applications, 219, 119681, 2023.
[24] Luo, F., et al. Innovative deep learning approach for cross-crop plant disease diagnosis. Computers and Electronics in Agriculture, 217, 2024.
[25] Barrett, H., and Rose, D. C. Perceptions of the Fourth Agricultural Revolution: what's in, what's out, and what consequences are anticipated? Sociologia Ruralis, 62, 162-189, 2022.