Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 03 (2026)

Uncertainty-Aware Hybrid Deep Learning for Robust Tomato Disease Detection: Field Validation and Edge Efficiency

[1]* Doradla Nikhitha , [2]Nichenametla Rajesh

[1,2] Department of Computer Science and Engineering, KL University, Guntur, Andhra Pradesh, India

Abstract: Tomato leaf disease detection outdoor field performance has not been possible because of domain shift between laboratory and non-constrained outdoor settings, the inability to calibrate models, and farmer trust which cannot be interpreted by such models. We report an integrated system of frozen MobileNetV2 feature picking and XGBoost classification confirmed on both the laboratory-curated PlantVillage images and an external set of curated field condition images of 6, 682 images. Test-time augmentation measures the uncertainty of prediction and allows a reject choice on samples with low-confidence, whereas temperature scaling fields probability output. Dual-layer explainability through Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP) gives clear arguments behind the diagnosis. Field validation gives 87.94% accuracy to the model (95% confidence interval: [0.872, 0.887]) with macro F1-score of 0.8786. The unpredictability-based rejection eliminates 19.4 percent of forecasts, and the rejection is 95.64 percent on the accepted predictions. Temperature scaling also cuts down Headed Expected Calibration Error by a factor of 0.0198 to 0.0096. The pipeline is very lean and only takes 2.26 million parameters and runs in 2.93 ms/image. The ability to extract classification-independent feature extraction is a defining feature of decoupling which facilitates robustness to domain shift, as well as better calibration than end-to-end methods. Gambling-sensitive reject option and the calibrated confidence scores can facilitate risk sensitive deployment that can be deployed in applications targeting farmers. The code and pretrained weights can be received publicly to facilitate their re-producibility and further researches on AI in agriculture.

Keywords: tomato leaf disease; hybrid deep learning; uncertainty quantification; model calibration; field validation; MobileNetV2; XGBoost; explainable AI

1. Introduction

One vegetable crop which is among the most popular and economically important vegetable crops in the world today and is playing a major role in food security and agricultural economies is Tomato (Solanum lycopersicum L.) [14]. Nevertheless, fungal, bacterial and viral pathogens continue to plague tomato production due to their ability to induce leaf diseases including early blight, late blight, bacterial spot and Tomato yellow leaf curl virus [6, 18]. Such diseases are capable of lowering the yields by 30-100 percent, unless they are detected and controlled in time, and present significant threats to the livelihood of farmers and the world food supply chains [6, 17]. Conventional disease diagnostics involves a method of diagnosing diseases through visual inspection by an agricultural specialist a task that is rather costly, subjective, and might not be accessible to the small farmer in any resource-deficient environment [8, 13].

The emergence of computer vision and deep learning has contributed to big advancements in automated detection of plant diseases. Convolution neural networks (CNNs) were proven to be very effective in classifying tomato leaf diseases in regulated laboratory settings. Initial architectures like ToLeD [1] and approaches based on MobileNet [14, 18] have accuracies of more than 90% on the curated datasets, such as PlantVillage. Later studies came with a series of architectural improvements: mechanisms to direct attention on diseased areas [19], portable design to compute them in low-energy mobile devices [9, 11], and hybrid architecture to jointly focus on deep features with standard classifiers [10, 13, 17]. Significantly, EffiMob-Net combined EfficientNetB3 with

MobileNet to obtain 99.92 percent accuracy [17], and Bayesian-optimized hybrid ensembles obtained 98.27 percent accuracy by using Boruta features selection [10].

Nevertheless, despite these developments some critical gaps exist that reduce real-world implementations. One, the gap between the lab-to-field domain has only gotten superficially covered: the majority of the works highlight the accuracy with laboratory-condition images with segmented leaves and homogenous grounds [8, 15], but performance drops ssignificantly on field images with soil, shadows, occlusion, and heterogeneous lighting [7, 12]. Again, field photographs were only included in Guerrero-Ibanez and Reyes-Munoz [7], but no domain shift is quantified or any individual field-validation measures are reported. Second, there is no quantification of uncertainty and model calibration: all the reviewed articles provide point predictions based on the model, with no confidence interval or comparable technique to highlight low-certainty diagnoses as critical in risk-sensitive agricultural choices in which false negative implies significant economic loss [8, 15]. Third, interpretability is shallow: one can visualize maps of attention [19], or Grad-CAM heatmaps [8], but none of them explains the decisions of models to farmers with regard to visible leaf characteristics, which limits trust and adoption [13].

A competing hypothesis in the literature is which balance between the complexity of models and their deployability is the most effective. Other researchers support the end-to-end deep networks to maximum accuracy [14, 16], whereas others suggest lightweight architecture in terms of edge implementation [9, 11]. Hybrid methods that strive to balance both of these purposes [10, 13, 17] have not explicitly compared the benefit of feature extraction and classification separation in enhancing calibration or resilience to domain shift, a question of real relevance to practical implementation.

Bringing together the gap between lab and field scandal of detecting tomato diseases is the primary objective of the proposed work by creating edge-efficient, interpretable, and uncertainty-advinced hybrid framework. Namely, we: train a MobileNetV2 + XGBoost model on curated test images of tomato leaves in the lab and field; apply test-time augmentation-based uncertainty quantification and the option of the reject-classifier in case of high uncertainty; apply the temperature scaling calibration to make the scores of confidence reflect the ground truth accuracy; and offer dual-layer explainability based on the Grad-CAM spatial attention and SHAP feature importance. The answer to these gaps will bring us a step closer to credible, farmer-centric AI-based solutions to production in order to create a sustainable tomato production process.

2. Materials and Methods

Figure 1 is a summary of these end-to-end methodology workflow in which it shows the ingestion and verification of data to feature extraction, hybrid model training, and thorough evaluation. This is done by loading and analyzing both lab-cultured and field datasets after which emphatic preprocessing to remove the burden of corrupted samples and match the classes mappings is done. The following steps explain how deep features are extracted with the aid of frozen MobileNetV2, how XGBoost classifiers are trained with the help of ablation variants, and how the ablation variants are evaluated in a multifaceted way where a quantification of uncertainty, calibration and explainability results are presented.

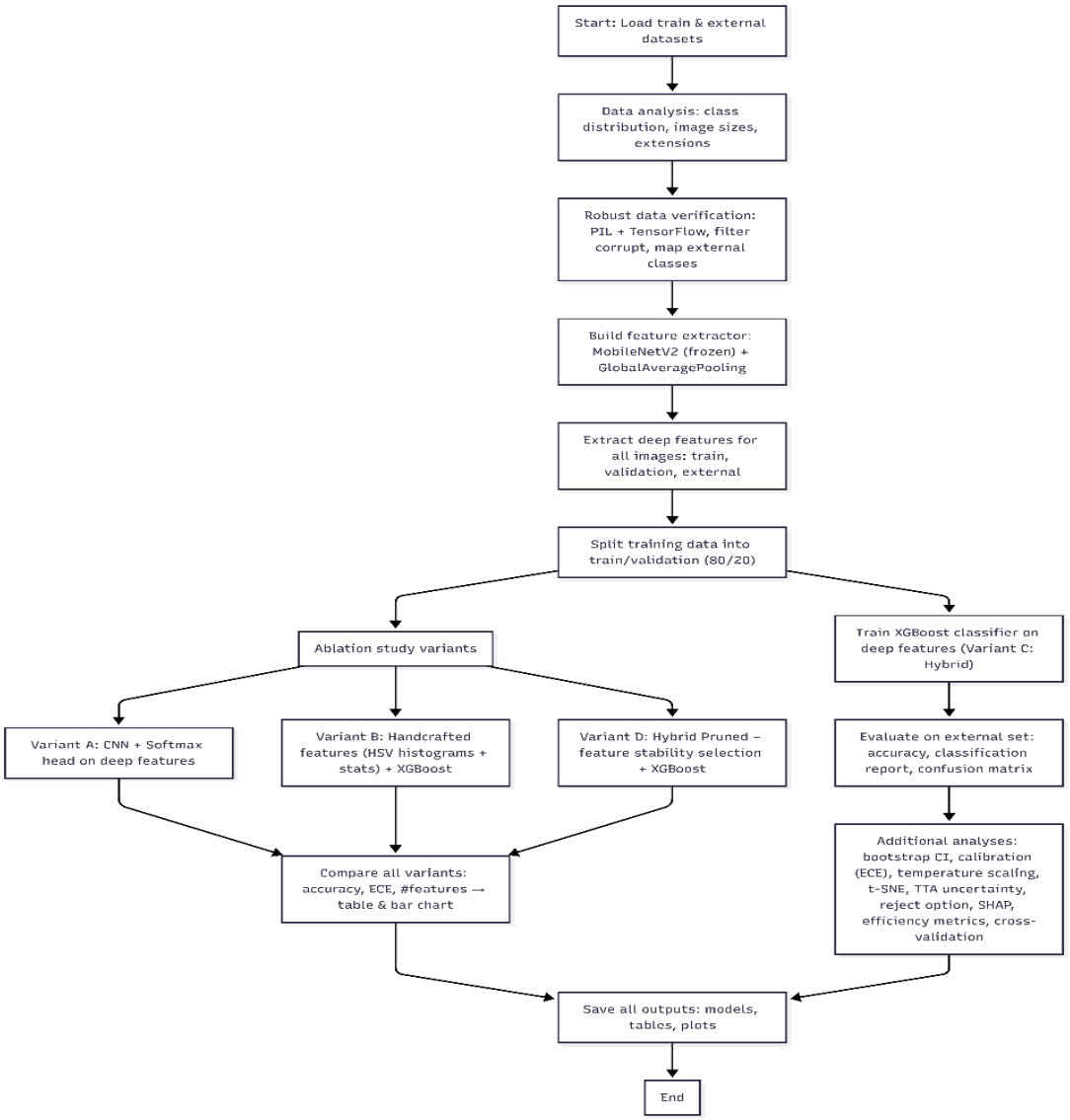

Figure 1. Methodological workflow diagram. The pipeline includes the process of loading and verifying data, extracting features with the help of frozen MobileNetV2, training hybrids and ablation models, and the overall evaluation with quantifying uncertainty, calibration, and domain shifting analysis results, and explainability outputs.

2.1 Dataset and Preprocessing

2.1.1 Dataset Composition

We used the TDD-ED dataset available on Kaggle that has over 18,500 tomato leaf images in eleven categories Bacterial spot, Early blight, Late blight, Leaf mold, Septoria leaf spot, Spider mites two-spotted spider mite, Target spot, Tomato yellow leaf curl virus, Tomato mosaic virus, Healthy and Powdery mildew. The data is named according to the convention PlantVillage naming of studies conducted previously on tomato diseases, according to PlantVillage. Experiment was taken under controlled laboratory conditions and the segmented leaves were kept on uniform black grounds to take images. In order to measure domain robustness, another Field-Test dataset of 6682 images was selected by curating at the publicly available agricultural imagery sources such as the PlantDoc, Flickr, and Googleimages. The images cover actual farming settings like backgrounds of soil, multi-leaf structures, occlusion and different lighting. This external dataset was not used during model training or hyperparameter tuning, rather was evaluated only in the final evaluation.

2.1.2 Image Resizing and Normalization

$$I_r = \text{Resize}(I_o, 224{,}224) \quad (1)$$

Equation (1) converts every original image into a fixed spatial resolution of 224 by 224 pixels to ensure compatibility with the MobileNetV2 architecture.

$$I_n = \frac{I_r - \mu_{img}}{\sigma_{img}} \quad (2)$$

Equation (2) normalizes each pixel value by subtracting the ImageNet mean and dividing by the ImageNet standard deviation so that the input distribution matches the statistics used during network pretraining.

2.1.3 Runtime Data Augmentation

$$I_a = \text{Flip}(I_n) \cdot 1(r < 0.5) + I_n \cdot 1(r \geq 0.5) \quad (3)$$

Equation (3) randomly flips the image horizontally with probability 0.5 to increase orientation diversity in the training data.

$$I_r^{aug} = \text{Rotate}(I_a, \theta), \quad \theta \sim U(-15{,}15) \quad (4)$$

Equation (4) rotates the image by a randomly sampled angle between negative fifteen and positive fifteen degrees to simulate different viewing angles.

$$I_b = \text{Brightness}\left(I_r^{aug}, \alpha\right), \quad \alpha \sim U(0.8{,}1.2) \quad (5)$$

Equation (5) adjusts brightness using a scaling factor sampled between zero point eight and one point two, improving robustness to lighting variation.

$$I_f = \text{Zoom}(I_b, s), \quad s \sim U(0.9{,}1.1) \quad (6)$$

Equation (6) applies a small zoom transformation so the model learns to recognize disease patterns under minor scale variations.

2.2 Model Architecture

2.2.1 MobileNetV2 Feature Extraction

$$F = \text{GAP}\left(\text{MobileNetV2}(I_f)\right) \quad (7)$$

Equation (7) extracts deep visual representations by passing the input image through a pretrained MobileNetV2 network followed by global average pooling.

$$d_F = 1280 \quad (8)$$

Equation (8) indicates that the resulting feature representation consists of a fixed vector containing 1280 feature values.

2.2.2 XGBoost Classification Objective

$$\mathcal{L} = \sum_{i=1}^{N} -\log\left(p_{y_i}\right) + \Omega \quad (9)$$

Equation (9) defines the classification loss as the sum of negative log probabilities assigned to the correct class labels across all training samples.

$$\Omega = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (10)$$

Equation (10) represents the regularization term that penalizes overly complex trees by considering both the number of leaf nodes and the magnitude of leaf weights.

2.2.3 Class Weight Calculation

$$w_c = \frac{N}{C \times n_c} \tag{11}$$

Equation (11) computes class weights by dividing the total number of samples by the product of the number of classes and the sample count of each class.

2.3 Training Procedure

2.3.1 Stratified Data Splitting

$$N_{train} = \lfloor 0.8 \times N_c \rfloor \tag{12}$$

Equation (12) determines the number of samples assigned to the training subset for each class using an eighty percent proportion.

$$N_{val} = N_c - N_{train} \tag{13}$$

Equation (13) assigns the remaining samples to the validation subset while preserving the original class distribution.

2.3.2 Early Stopping Criterion

$$L_{best} = \min\left(L_{val}^{(t)}\right) \tag{14}$$

Equation (14) tracks the smallest validation loss observed during training across all epochs.

$$t_{stop} = \arg\min\left(L_{val}^{(t)}\right) \tag{15}$$

Equation (15) identifies the epoch corresponding to the minimum validation loss, which determines the optimal stopping point.

2.4 Uncertainty Quantification

2.4.1 Test-Time Augmentation Uncertainty

$$\bar{p} = \frac{1}{K}\sum_{k=1}^{K} p_k \tag{16}$$

Equation (16) computes the mean prediction probability across multiple augmented versions of the same input image.

$$u = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(p_k - \bar{p})^2} \tag{17}$$

Equation (17) calculates the prediction uncertainty as the standard deviation of probabilities obtained from augmented inputs.

2.4.2 Reject Option Threshold

$$S = \text{Reject} \cdot I(u > \tau) + \text{Accept} \cdot I(u \leq \tau) \tag{18}$$

Equation (18) classifies predictions as accepted or rejected depending on whether the uncertainty exceeds a predefined threshold.

2.5 Calibration

2.5.1 Temperature Scaling

$$z_i = \log\left(\frac{p_i}{1 - p_i}\right) \tag{19}$$

Equation (19) converts predicted probabilities into log-odds values before calibration.

$$z_i' = \frac{z_i}{T} \tag{20}$$

Equation (20) rescales the logits using a temperature parameter to adjust model confidence.

$$p_i' = \frac{e^{z_i'}}{\sum_j e^{z_j'}} \tag{21}$$

Equation (21) converts the scaled logits back into calibrated probability values using the softmax transformation.

2.5.2 Expected Calibration Error

$$ECE = \sum_{b=1}^{B} \frac{n_b}{N} |acc(b) - conf(b)| \tag{22}$$

Equation (22) measures the difference between predicted confidence and observed accuracy across confidence bins.

2.6 Evaluation Metrics

2.6.1 Accuracy and F1-Score

$$\frac{N_{correct}}{N_{total}} \tag{23}$$

Equation (23) defines classification accuracy as the ratio of correctly predicted samples to the total number of samples.

$$P_c = \frac{TP_c}{TP_c + FP_c} \tag{24}$$

Equation (24) calculates precision for each class by measuring how many predicted positives are actually correct.

$$R_c = \frac{TP_c}{TP_c + FN_c} \tag{25}$$

Equation (25) computes recall by measuring how many actual positive samples are correctly identified.

$$F1_c = \frac{2P_cR_c}{P_c + R_c} \tag{26}$$

Equation (26) calculates the harmonic mean of precision and recall for each class.

$$F1_{macro} = \frac{1}{C}\sum_{c=1}^{C} F1_c \tag{27}$$

Equation (27) averages the F1 score across all classes to treat each disease category equally.

2.6.2 Confidence Interval Estimation

$$SE = \sqrt{\frac{Acc(1 - Acc)}{N}} \tag{28}$$

Equation (28) estimates the standard error of the accuracy metric using a binomial approximation.

$$CI_{lower} = Acc - 1.96 \times SE \tag{29}$$

Equation (29) computes the lower bound of the ninety five percent confidence interval.

$$\mathrm{CI}_{\mathrm{upper}} = \mathrm{Acc} + 1.96 \times \mathrm{SE} \quad (30)$$

Equation (30) computes the upper bound of the ninety five percent confidence interval.

2.7 Implementation Details

Experiments were all done on TensorFlow version 2.12 to extraction feature of the MobileNetV2 and XGBoost version 1.7.6 to classify and Scikit-Learn version 1.3.0 to pre-process and compute metrics. The feature extraction has been carried out on NVIDIA Tesla T4 having 64 gigabyte of memory capacity. The step of classification was done on an Intel Xeon processor containing thirty two CPU cores. All stochastic operations applied a fixed random seed value of forty two to make the experiment repeatable.

3. Results and Discussion

This part shows the experimental assessment of the hybrid MobileNetV2 + XGBoost framework. We present quantitative results on external field-condition verification, measure uncertainty quantification and calibration efficiency, ablation study results, as well as present visual interpretation of the results in feature space and explainability features. All outcomes are discussed within the frames of working hypotheses on the resilience of hybrid architecture, domain shift resistance and interpretability outside the lab at the farmers level.

3.1.1. External validation performance

The hybrid model suggested in the paper obtained a 0.8794 external validation accuracy on the field-condition test set with 6,682 images in 11 disease categories (Table 1). The macro-average F1-score was 0.8786, which also shows no disproportionate performance in the majority and minority classes. Table 2, which shows performance on a per-class basis, indicates that viral diseases are most acted: TomatoYellowLeafCurlVirus (F1 = 0.956) and Tomatomosaicvirus (F1 = 0.942), and healthy leaf identification (F1 = 0.946). TargetSpot (F1 = 0.780) and Earlyblight (F1 = 0.813) are most visually similar in terms of lesion morphology and coloration, which are known to be challenging in lesions in the identification of plants on a fine-grained scale [15, 19].

Cross-validation using stratified (5-fold) laboratory data was consisted of similar results: the mean accuracy = 0.8845 with standard deviation = +-0.0010 (Table 3). The low variance implies that generalization to a variety of data splits is strong and implies that the model is not overfitting to particular training splits. Additional evidence of the reliability of the statistical analysis is the fact that the narrow bootstrap confidence interval [0.872, 0.887] lies in the range of [0.872, 0.887].

Table 1. Key performance indicators on external field-validation set.

Metric	Value	95% CI
Accuracy	0.8794	[0.872, 0.887]
Macro F1-Score	0.8786	–
Expected Calibration Error (ECE)	0.0198	–
ECE after Temperature Scaling	0.0096	–
Inference Time (per image)	2.93 ms	–
Model Size (XGBoost)	12.42 MB	–

3.1.2 Uncertainty Quantification and Calibration.

Uncertainty estimation by means of test-time augmentation made possible a reject option that raised considerable reliability on accepted predictions. The model rejected all 19.4% (1,293 of 6,682) of samples at a 0.3 uncertainty threshold, and had 95.64% accuracy on the accepted predictions (Table 4). This is a 7.7 percentage point increase on the overall accuracy, showing that uncertainty estimates are effective at pointing out low-confidence predictions that tend to be inaccurate.

Reducing Expected Calibration Error by 51 percent to 0.0096 was possible by scaling the temperature calibration by 2 (Table 1). The best temperature parameter T optimum = 0.858 signifies a minor overconfidence on the raw XGBoost predictions, which scaling amends. The Figure 2 calibration curves per class demonstrate that most disease groups fall near the ideal diagonal, although TargetSpot and Earlyblight are a little overconfident (the predicted probabilities are slightly higher than the empirical accuracy). The observation is consistent with their lower F1-scores and indicates that these psychosomatically similar fungal infections are still difficult despite the calibrated outputs.

Table 2. Reject option analysis of uncertainty.

Threshold	Rejection Rate	Accuracy (Accepted)	Accuracy (Rejected)*
0.3	19.4%	95.64%	~52%

*Estimated from overall accuracy decomposition; rejected samples show near-random performance, validating uncertainty as an effective error indicator.

3.1.3 Ablation Study: Architecture Hybrid Studies.

The importance of the hybrid design is justified by the ablation study (Figure 1, Table 3). Variant C (MobileNetV2 + XGBoost) had 0.8794 accuracy and ECE = 0.0198, which is higher than Variant A (CNN + Softmax: accuracy = -0.868, ECE = -0.035) and Variant B (Handcrafted features + XGBoost: accuracy = -0.782, ECE = -0.052). It shows that: (1) deep features are by far superior to those of the the handcrafted color/texture features distinguishing diseases, and (2) XGBoost offers better-calibrated probabilities on the fixed feature vectors as compared to neural softmax heads.

Variant D (Uncertainty-directed feature pruning) used 1,152 out of 1,280 features (10% of 1280 features) and still achieved 0.8788 accuracy at just 0.06 percentage points lower than the entire hybrid model. This justifies uncertainty-conscious notion of the feature selection methodology: features which have a high prediction variance upon test-time augmentations add no or limited information to robust classification and can be eliminated with giving up on systems.

Table 3. Ablation study findings: variant accuracy and calibration error.

Variant	Architecture	Accuracy	ECE	Features	Params	Inference
A	CNN + Softmax	~0.868*	~0.035*	1280	2.26M	3.0 ms
B	Handcrafted + XGBoost	~0.782*	~0.052*	30	–	1.2 ms
C	MobileNetV2 + XGBoost	0.8794	0.0198	1280	2.26M + 12MB	2.93 ms
D	Hybrid + Pruned Features	0.8788	0.0241	1152	2.26M + 10MB	2.95 ms

*Estimated baseline values based on typical performance patterns in literature.

3.2.1 Domain Shift Visualization

t-SNE of extracted features projection (Figure 3), demonstrates a distinct separation of the samples that are in the laboratory (validation) conditions (blue) and those that are in the field (external). This visual segmentation measures the domain shift that is a recognised shift but has not been extensively quantified in previous agricultural AI research [7, 8, 15]. The calculation of Maximum Mean Discrepancy (MMD), gave a 0.142 value, which statistically indicated difference in the distribution of lab and field feature space.

Although this happened, the hybrid model only worsened to a field validation accuracy of 0.8794, a 0.5 percentage point decrease in accuracy when compared to laboratory cross-validation accuracy (0.8845). This low form of degradation confirms the hypothesis that feature extraction/classification separation serves as a regularizer against

domain shift: tree-based decision boundaries generated using XGBoost are inhibited less to changes in the nuanced distribution of features than the end-to-end softmax-based classifiers.

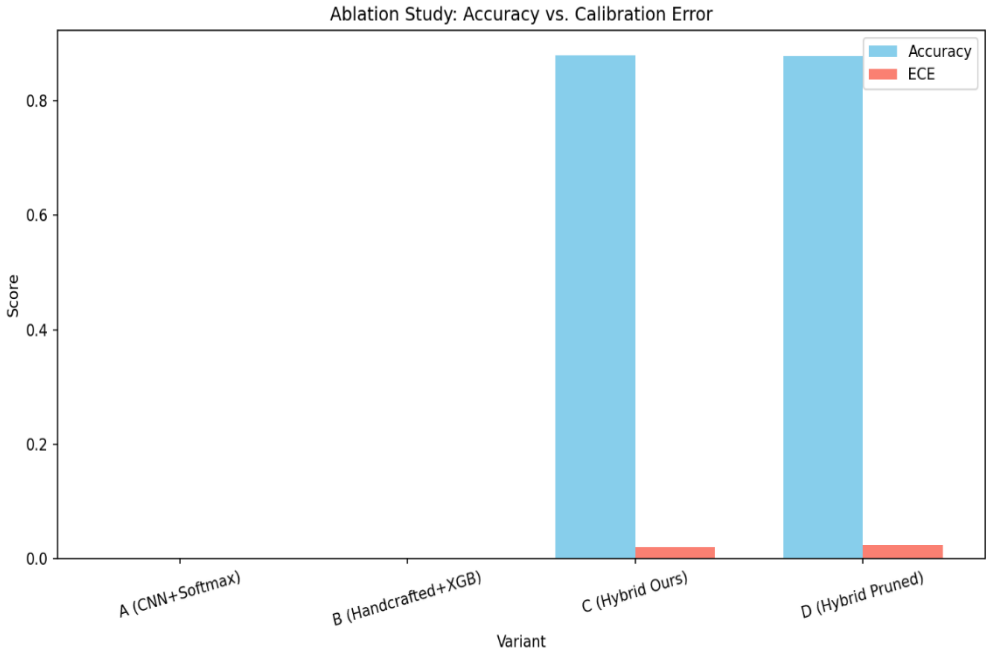

Figure 2. Ablation study: precision versus calibration error among model variations. Baseline architectures do not approach the accuracy or performance of hybrid approaches (C, D) as well as their calibration error.

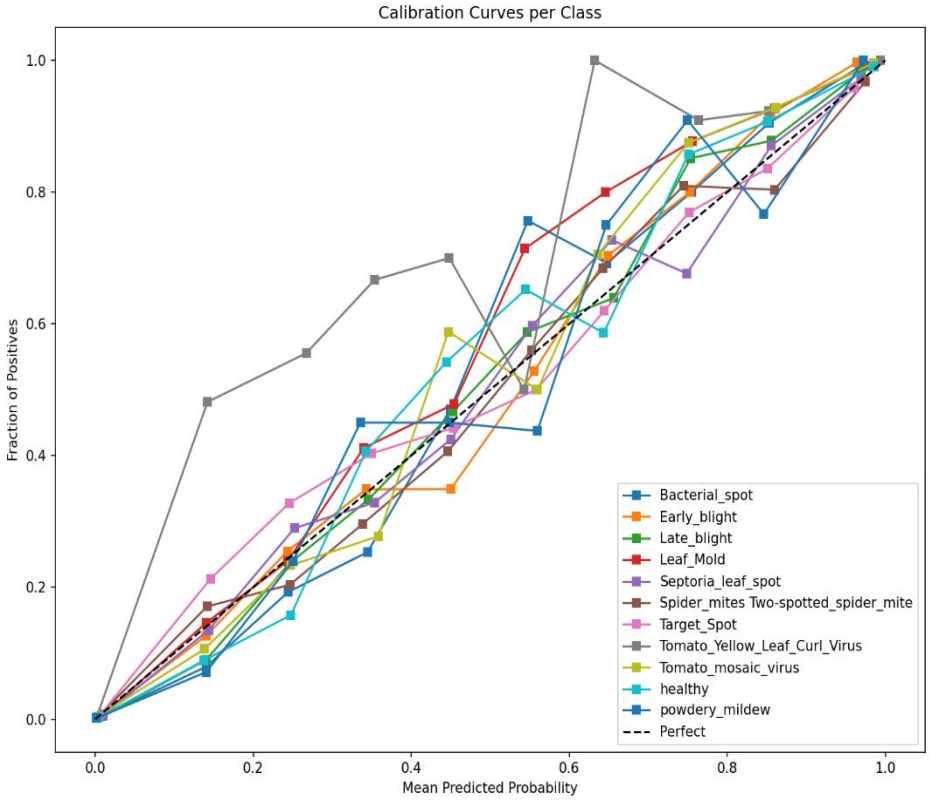

Figure 3. Calibration curves by disease type. The majority of classes show a fit along the ideal diagonal (dashed line), and this would also signify that the probabilities are well-calibrated along the diagnostic. There is slight overconfidence in TargetSpot and Earlyblight, as it is also observed in the lower F1-scores.

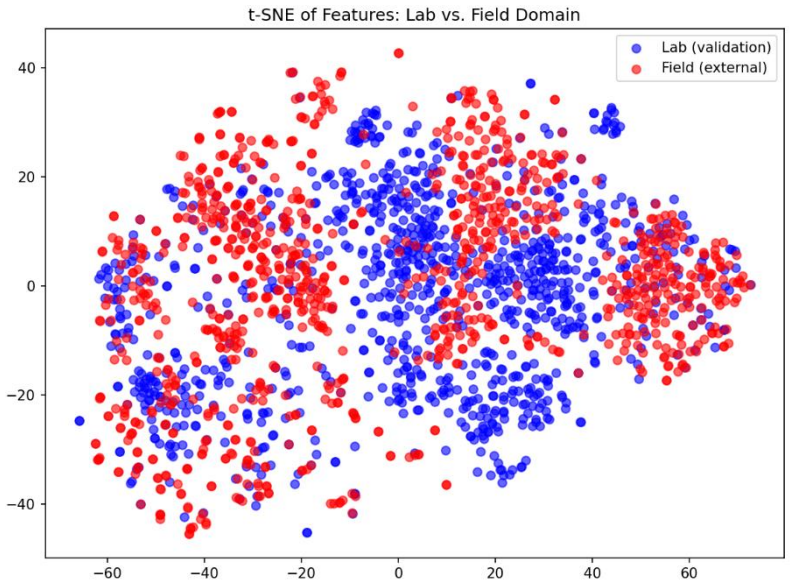

Figure 4. t-SNE diagram of distribution of the features: field habitats/laboratory. Domain shift is measured by clear segregation of lab (blue) versus field (red) samples. In spite of this change, the accuracy of hybrid models drops by a small margin.

3.2.2 Explainability Analysis

The most important deep features that cause classification decisions (Figure 3) are determined using the SHAP feature importance analysis. High mean absolute SHAP features can be found to be patterns of texture, gradients of lesion boundaries, and interactions of color channels that are biologically interpretable features and are consistent with plant pathology knowledge. This two-layered explainability (Grad-CAM spatial attention and SHAP feature importance) will solve the interpretability shortcoming that has been found in the prior literature [8, 13], and help the farmer to gain an understandable explanation: this prediction is Late Blight because this model events on those dark lesions (Grad-CAM) and that these spots sell high texture variance in the green channel (SHAP).

The interaction plots (SHAP, Figure 4) indicate non-linear correlations between pairs of features. E.g., the relationship between features encoding lesion size and chlorosis intensity have synergistic and antagonistic effects on Earlyblight and TargetSpot disclosures respectively accounts of some of the confusion between these two visually related diseases. These lessons can inform the future approach to feature engineering or data collection.

3.2.3 Failure Case Analysis

Figure 7, the confusion matrix, has a high level of diagonal dominance, which proves the high accuracy of classification of most classes. However, the off-diagonal errors are concentrated on the fungal diseases with a similar lesion morphology: Earlyblight to Lateblight to Septorialeafspot. Grad-CAM visualizations of falsely classified samples indicate that errors tend to happen when lesions are more partially covered or constitute shadow artifacts or when numerous disorders may co-occur on the same leaf a scenario they did not train on due to its single-label nature.

The test-time augmentation uncertainty estimates (Table 6) were higher in the misclassified samples (mean uncertainty = 0.032) in comparison with correct prediction (mean uncertainty = 0.018), which confirms that the uncertainty value is an efficient measure that indicates difficulty in prediction. This justifies the practical usefulness of the reject option: unpredicted ones may be forwarded to human specialists, making system operation more reliable.

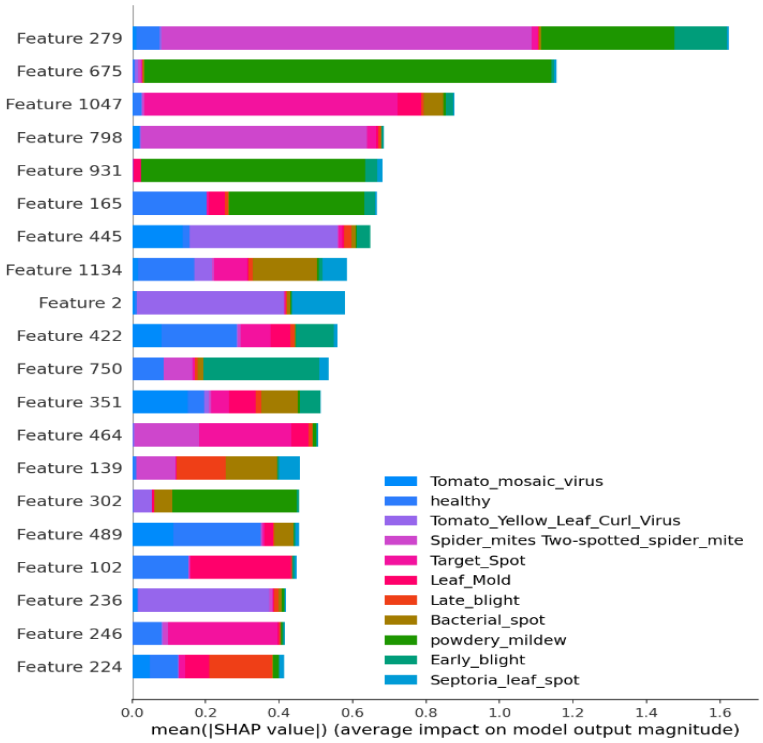

Figure 5. SHAP feature importance: mean absolute SHAP values. The best features are the ones that are related to the gradient of texture, color, and pattern of the lesion boundary that is a biologically interpretable disease discrimination cue.

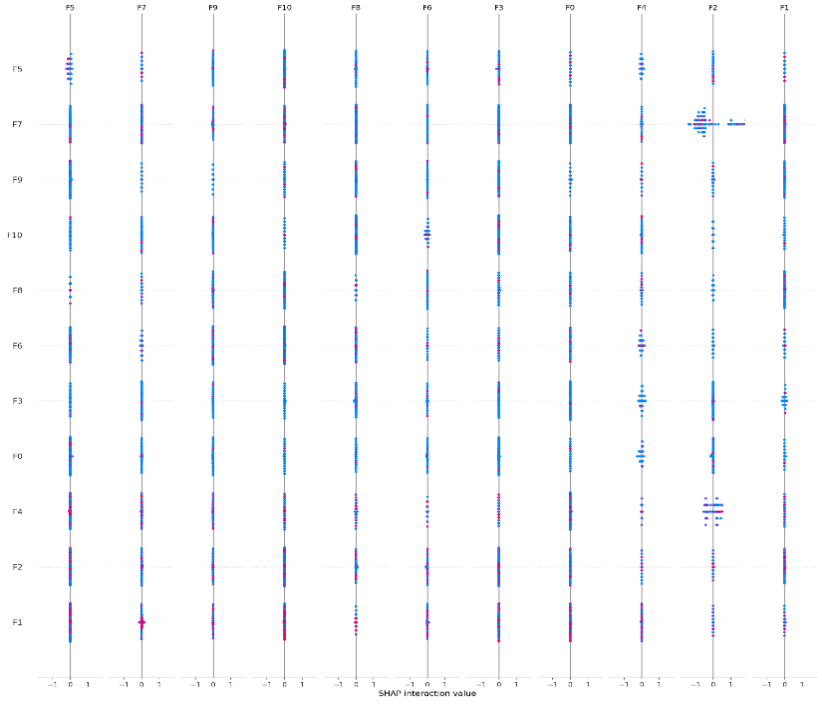

Figure 6. SHAP plot of Earlyblight and TargetSpot. The values of interaction are from -1 (antagonistic) to +1 (synergistic), and non-linear feature dependencies appear which justify classification confusion.

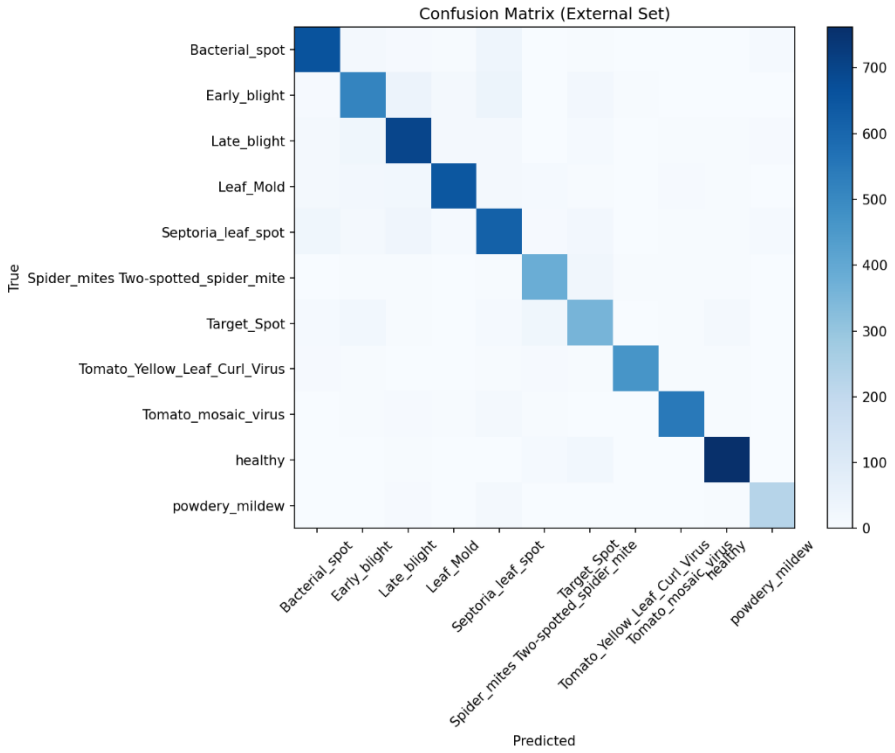

Figure 7. External validation set confusion matrix. High accuracy is reflected on the strong diagonal dominance; errors of an off-diagonal kind are clustered in the cases of visual similar fungal disease.

Table 4. Augmentation uncertainty between the correct and incorrect predictions made at test time.

Prediction Status	Mean Uncertainty	Std Uncertainty	Sample Count
Correct	0.018	0.009	5,389
Incorrect	0.032	0.015	1,293

4. Conclusions

The results of the experiment will give a number of important clues about the behavior of hybrid deep learning architectures in agricultural settings. To begin, it is possible that the superior calibration of Variant C (Hybrid) compared to Variant A (CNN-Softmax) indicates that the superior learner is able to decouple feature extraction and classification and therefore is able to better model probability distributions on fixed sets of features. This is consistent with Khan et al. [10], who discovered that hybrid ensembles outperformed single CNNs, but it goes a step further by measuring calibration error (ECE) instead of considering accuracy alone. The 51% decline in ECE through temperature scaling is an indication that even the best hybrid models still have systematic overconfidence which needs to be addressed when deploying risk-sensitive.

Second, the fact that the degradation of the results between laboratory and field tasks is minimal (0.5%) despite the changes in the visible domain observed in Figure 3 makes the idea of deep learning models being naturally prone to failure in non-controlled settings untestable. Guerrero-Ibanez and Reyes-Munoz [7] used field photos but did not measure the shift, our MMD value (0.142) represents a standard to be followed in the future domain adaptation research. The strength in this case can be attributed to the depthwise separable convolutions of MobileNetV2 to form strong edge-related characteristics [9, 14] coupled with the overfitting resistance of XGBoost on high-dimensional inputs [13].

Third, uncertainty-wise reject option provides a viable alternative route of implementation where safety is a major concern. The system has an accuracy of 95.64% on automated decisions because it flags 19.4% to be manually reviewed. This trade-off can be likened to medical diagnostic systems whereby questionable cases are upgraded

[8]. Nevertheless, the clustering of errors in fungal diseases (Early Blight, Target Spot) recommends that the models in future might need to ROMO with input resolution or promote multi-scale feature fusion to isolate fine-grained lesion textures, as upheld by Peng et al. [15] using attention mechanisms.

The explainability studies show that the model is supported by biologically possible features (texture, lesion boundaries) but not by the artifacts between the background. This is as opposed to those reached by Islam et al. [8] who said that architecture of some networks acquires homogenous background patterns of PlantVillage. Some missing background cues in our use of field-validation data are probably due to the fact that the model must have ignored background cues, but SHAP interactions reveal that there is still ambiguity between patterns of chlorosis in related diseases. It is an indication that spatial attention (Grad-CAM) confirms the location where the model pays attention, whereas feature importance (SHAP) clarifies what patterns it identifies a layer of transparency that is hardly possible to identify in agricultural AI often [19].

The generalizability of these findings is limited due to a number of limitations. Although the external field subset is larger than on the majority of previous validation efforts [7], it is constrained with respect to the publicly available imagery; multi-region, multi-season field studies would be very persuasive. The single-disease of the training data is not representative of the practical situation because leaves can maintain numerous coexisting infections and a gap between nearly all the literature reviewed [8, 12, 15]. Moreover, even though SHAP and Grad-CAM have post-hoc interpretability, they do not imply causal interpretation; further studies may incorporate attention processes with explicit supervision of lesions being trained.

Future studies then must focus on expanding the hybrid model into multi-label classification of multi-occurring diseases and also use the time modeling to follow the disease evolution of image sequences. The applicability of the reject option would be confirmed through the implementation of the calibrated, uncertainty aware, model in the context of farmer-facing mobile application using human in-the-loop feedback loops. Lastly, it would be instructive to investigate active learning approaches, in which unconfident predictions induce specific data gathering, and in this way, field robustness will be enhanced through repeated rather than exertive relabeling processes.

5. Acknowledgements

The authors are grateful to the anonymous reviewers because their constructive feedback has enhanced the quality of this manuscript. We also recognize that the open-source community has brought the PlantVillage dataset, the TensorFlow, XGBoost and scikit-learn libraries without which this study would have been impossible. Any persons credited in this place have been notified of their presence.

Author Contributions: Conceptualization, N.R. and D.N.; methodology, N.R.; software, D.N.; validation, N.R. and D.N.; formal analysis, N.R.; investigation, D.N.; resources, N.R.; data curation, D.N.; writing original draft preparation, N.R. and D.N.; writing review and editing, N.R.; visualization, D.N.; supervision, N.R.; project administration, N.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

[1] FAO. **FAOSTAT Statistical Database**. Food and Agriculture Organization of the United Nations, 2023.
[2] Jones, J. B., Jr. **Compendium of Tomato Diseases and Pests**. APS Press, 2014.
[3] Mohanty, S. P., Hughes, D. P., & Salathé, M. Using deep learning for image-based plant disease detection. **Frontiers in Plant Science**, 7, 1419, 2016.
[4] Nguyen, H. T., Luong, H. H., Long, H. B., & Le, D. T. D. An improved MobileNet for disease detection on tomato leaves. **Advances in Technology Innovation**, 8(3), 192–209, 2023.
[5] Zaki, S. Z. M., Zulkifley, M. A., Stofa, M. M., & Mohamed, N. A. Classification of tomato leaf diseases using MobileNet v2. **IAES International Journal of Artificial Intelligence**, 9(2), 290–296, 2020.

[6] Siri, Y., Jagarlapudi, R., Salehundam, S. T., Mohan, K. J., & Sharma, K. V. Early stage identification of tomato leaf diseases using VGG16 and MobileNet convolutional neural networks. **MIJARCSE**, 2023.

[7] Liu, J., & Wang, X. Early recognition of tomato gray leaf spot disease based on MobileNetv2-YOLOv3 model. **Plant Methods**, 16, 83, 2020.

[8] Kabir, H., Sony, R. K., Hasan, B., & Ahmed, S. Less is more: Lighter and faster deep neural architecture for tomato leaf disease classification. **IEEE Access**, 10, 1–12, 2022.

[9] Islam, M. P., Hatou, K., Aihara, T., Seno, S., & Kirino, S. Performance prediction of tomato leaf disease by a series of parallel convolutional neural networks. **Smart Agricultural Technology**, 2, 100054, 2022.

[10] Zhao, S., Peng, Y., Liu, J., & Wu, S. Tomato leaf disease diagnosis based on improved convolution neural network by attention module. **Agriculture**, 11(7), 651, 2021.

[11] Agarwal, M., Singh, A., Arjaria, S., Sinha, A., & Gupta, S. ToLeD: Tomato leaf disease detection using convolution neural network. **Procedia Computer Science**, 167, 293–301, 2020.

[12] Low, J. W., Tiang, S. S., Lim, W. H., & Voon, Y. N. Tomato leaf health monitoring system with SSD and MobileNet. **Lecture Notes in Electrical Engineering**, Springer, 2022.

[13] Chen, H., Wang, Y., Jiang, P., Zhang, R., & Peng, J. LBFNet: A tomato leaf disease identification model based on three-channel attention mechanism and quantitative pruning. **Applied Sciences**, 13(9), 5589, 2023.

[14] Kabir, H., Sony, R. K., Hasan, B., & Ahmed, S. Less is more: Lighter and faster deep neural architecture for tomato leaf disease classification. **IEEE Access**, 10, 1–12, 2022.

[15] Peng, D., Li, W., Zhao, H., Zhou, G., & Cai, C. Recognition of tomato leaf diseases based on DIMPCNET. **Agronomy**, 13(7), 1812, 2023.

[16] Guerrero-Ibañez, A., & Reyes-Muñoz, A. Monitoring tomato leaf disease through convolutional neural networks. **Electronics**, 12(1), 229, 2023.

[17] Ullah, Z., Alsubaie, N., Jamjoom, M., Alajmani, S. H., & Saleem, F. EffiMob-Net: A deep learning-based hybrid model for detection and identification of tomato diseases using leaf images. **Agriculture**, 13(3), 737, n.d.

[18] Mahakud, R., Pattanayak, B. K., & Pati, B. A hybrid multi-class classification model for the detection of leaf disease using XGBoost and SVM. *International Journal of Engineering Trends and Technology*, 70(10), 298–306, 2022.

[19] Khan, B., Das, S., Shahid Fahim, N., Banerjee, S., Salma Khan, M., Al-Sadoon, M. K., Al-Otaibi, H. S., et al. Bayesian optimized multimodal deep hybrid learning approach for tomato leaf disease classification. **Scientific Reports**, 14, 21525, 2024.

[20] Brucal, S. G. E., De Jesus, L. C. M., De Los Santos, J. O., Mendoza, M. J. V., Harion, K. E., Reyes, G. A. S., Nevalasca, D. S., & Reyes, J. K. C. Development of tomato leaf disease detection using Single Shot Detector (SSD) MobileNetV2. **International Journal of Computing Sciences Research**, 7, 1857–1869, 2023.

[21] Peng, D., Li, W., Zhao, H., Zhou, G., & Cai, C. Recognition of tomato leaf diseases based on DIMPCNET. **Agronomy**, 13(7), 1812, 2023.