TOWARDS TRUSTWORTHY EXPLAINABLE AI FOR TOMATO LEAF CLASSIFICATION: A FIDELITY AND STABILITY ANALYSIS OF GRAD-CAM AND LIME

Abstract:

Deep learning models achieve state-of-the-art performance in automatic plant disease classification; however, their decision opacity remains a major barrier to adoption in high-stakes agricultural applications. Indeed, trust from agronomy experts depends directly on the ability of these systems to provide reliable and interpretable justifications for their predictions. In this work, we present an evaluation of two post-hoc explainability methods, Grad-CAM and LIME, applied to an EfficientNetV2B3 model trained on a tomato leaf image dataset. The analysis combines a qualitative assessment of activation maps and explanatory masks with a quantitative evaluation based on DAUC and IAUC metrics, complemented by a stability analysis under input perturbations.

The results reveal a trade-off between fidelity and stability of explanations. LIME achieves higher fidelity according to insertion and deletion metrics (IAUC = 0.9268, DAUC = 0.4187), whereas Grad-CAM provides more stable explanations (0.0002 vs. 0.0025). Qualitative analyses further show that both methods highlight disease-relevant agronomic symptoms. These findings underscore the complementarity of Grad-CAM and LIME for explainability. Grad-CAM facilitates the localization of disease-relevant regions, whereas LIME provides a more detailed characterization of the influential pixels within those regions. Together, they offer a coherent and complementary explanatory framework that strengthens the overall explainability of the model.

Key words:-

XAI; Deep Learning; Grad-CAM; LIME; Plant Disease Detection; Fidelity; Stability EfficientNet

…….....

Introduction:-

In our previous study**[28],**we introduced a Convolutional Neural Network (CNN) to classify tomato leaf diseases, achieving an overall accuracy of 94%. These results highlight the strong potential of deep learning methods to improve decision-support systems in precision agriculture, particularly for plant disease diagnosis.

However, despite their high performance, CNN models present a major limitation related to decision opacity, especially when dealing with data containing epistemic uncertainty. A critical question arises at each inference step: what internal mechanisms enabled the model to generate a specific prediction?. This issue lies at the core of the current limitations of deep learning architectures, which are commonly considered as “black-box” models. While these models demonstrate excellent performance in terms of accuracy and generalization, their lack of interpretability represents a significant barrier to their adoption in sensitive domains, where it is essential to understand and justify algorithmic decisionsTo address these challenges, the field of Explainable Artificial Intelligence (XAI) has emerged, based on two key concepts: interpretability and explainability. As highlighted by Roscher et al. (2020)**[1]** and Fuhrman et al. (2022)**[2]**, interpretability refers to the ability to translate an abstract concept derived from a model into a human-understandable representation, which is the case for models such as linear regression or decision trees. Explainability, on the other hand, is a more demanding concept that requires not only interpretation but also additional contextual information, particularly for complex models such as CNNs or Transformers.The agricultural sector is particularly affected by these challenges. Critical decisions such as targeted pesticide spraying or isolating infected crops, when based on non-interpretable predictions, may reduce farmers’ trust and slow down the adoption of intelligent agricultural systems.

In this context, the objective of this work is to propose an explainability approach applied to a CNN model based on EfficientNet, trained using fine-tuning, for tomato leaf disease detection. We use the Gradient-weighted Class Activation Mapping (Grad-CAM) method to generate activation maps highlighting the spatial regions that contribute most to the prediction, as well as the Local Interpretable Model-agnostic Explanations (LIME) method to locally identify the pixels or features influencing the model’s decision.

UNDER PEER REVIEW IN IJAR

The main objective is to demonstrate that the combination of Grad-CAM and LIME provides more complete and complementary explanations of the model predictions, thereby improving the understanding of the model's decision-making mechanisms.

Related Work:-

The explainability of deep learning models is essential for building user trust and facilitating the adoption of these systems, as discussed by Melkamu Mersha et al. (2025)**[3]**Among the commonly used approaches are perturbation-based methods (LIME, SHAP, CFE) and gradient-based methods (saliency maps, CAM, LRP).

These different technical approaches are applied in several domains, particularly in computer vision, as shown in the work of Nguyen Van Tu et al. (2025)**[4]**, which presents the applicability, advantages, and limitations of methods such as Saliency Maps, Concept Bottleneck Models (CBM), prototype-based approaches, and hybrid methods.Practical applications of explainability have also been proposed in other domains, notably in medical imaging, where Apoorva Aravindkumar et al. (2026)**[5]**developed methods to make deep learning models explainable while maintaining their usability in clinical environments. Similar applications are also emerging in agriculture, particularly in computer vision tasks related to plant disease detection.

More specifically, EfficientNet architectures have been the subject of dedicated explainability studies. For example, Jyothi Peta et al. (2024)**[6]** combined several techniques such as Grad-CAM, LIME, and SHAP to interpret the predictions of models applied to breast cancer imaging.

All domains are concerned with the issue of explainability, and the agricultural domain is no exception, particularly for leaf disease classification. Kaihua Wei et al. (2022) **[7]**, for instance, used classical architectures such as VGG, GoogleNet, and ResNet for fruit disease detection using the PlantVillage dataset and proposed a local explainability analysis using methods such as SmoothGrad, LIME, and Grad-CAM.

Unlike these studies, which rely on PlantVillage images acquired under laboratory conditions and on architectures such as VGG, GoogleNet, or ResNet, our study focuses on the explainability of a tomato leaf disease classification model using images captured under real field conditions. While combinations of LIME and Grad-CAM are still being explored and the EfficientNetV2-B3 architecture remains relatively underexplored in real agricultural contexts, we propose to use this black-box model and leverage the combination of Grad-CAM and LIME to provide more complete and complementary explanations of the model's decisions. The next section details the methodology used.

Proposed Methodology:-

In this study, we propose a methodology for tomato leaf disease classification using EfficientNetV2-B3 combined with visual explanation techniques. The overall workflow is illustrated in the figure below. The main steps are as follows:

- **Dataset and Model Summary:** The test dataset is prepared by collecting tomato leaf images in real-world conditions. The EfficientNetV2-B3 model, pre-trained on ImageNet, is fine-tuned on this dataset to perform multi-class classification of leaf diseases.
- **Image Processing for Inference:**Each image undergoes resizing, normalization, and optional data augmentation to ensure consistency and improve model generalization.
- **Inference:**The fine-tuned model predicts the class of each input image, producing probability scores for all target classes.
- **Explainable AI Techniques:**Visual explanations are generated using Grad-CAM to highlight spatial regions contributing most to the model's decision, and LIME to identify influential pixels or superpixels on a per-instance basis.
- **Evaluation:**The quality of explanations is assessed qualitatively through visual inspection of heatmaps and mask, and quantitatively using metrics such as fidelity and stability between highlighted regions and annotated disease areas.

This methodology ensures both high-performance classification and explainable predictions, providing insights into the model's decision-making process for real-world agricultural applications.

UNDER PEER REVIEW IN IJAR

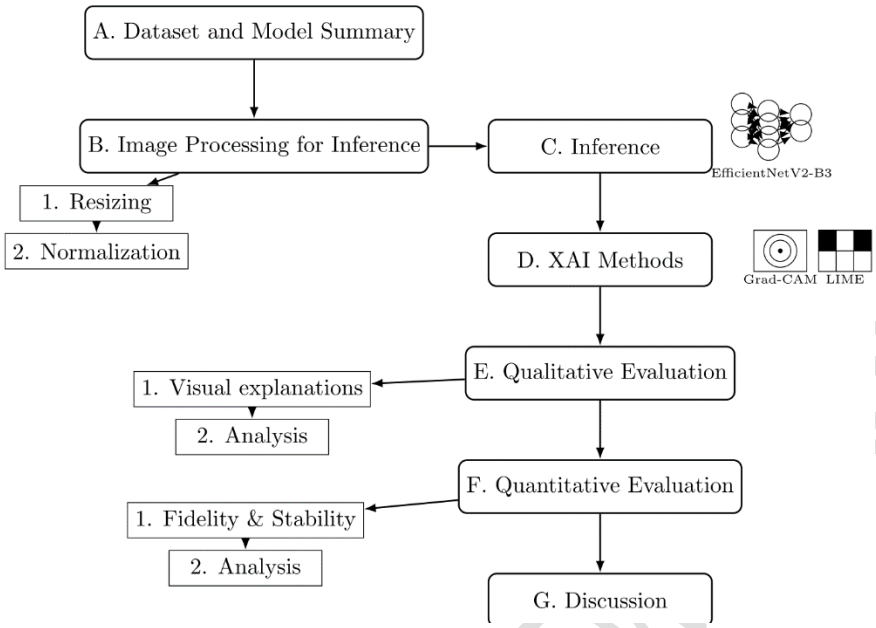


Figure 1: Methodology

A. Dataset and Model Summary:-

Dataset:-

The dataset used in this study consists of 1,831 images of tomato leaves (*Solanum lycopersicum*), collected in Mboro, Thiès region (Senegal), the hot dry season. It includes three classes: Tomato Yellow Leaf Curl Virus (n = 611), Early Blight (n = 604), and Healthy leaves (n = 614). The images were captured under varying lighting, angle, and background conditions, reflecting real-world agricultural constraints.

The dataset was partitioned into training, validation, and test subsets using a stratified sampling strategy, ensuring the preservation of class distributions across all splits. The allocation was performed according to fixed proportions (80%/10% / 10%), leading to exact class-wise distributions as reported in Table 1.

Table 1 : Dataset distribution by class and subset

Class	Total	Training	Validation	Test
TYLCV	611	499	61	51
Early Blight	604	492	61	51
Healthy	614	499	63	52
Total	**1829**	**1490**	**185**	**154**

Model:

The proposed model is based on the fine-tuning of the EfficientNetV2-B3 architecture initialized with ImageNet pre-trained weights. The last four layers of the network were unfrozen and retrained on the target dataset in order to adapt high-level features to the tomato leaf disease classification task. In addition, three fully connected dense layers were added before the final classification layer to improve task-specific feature learning and adapt the model to the multi-class classification problem.

After the fine-tuning process, the model achieved an overall classification accuracy of 94% on part the test dataset. The classification performance on a portion of the test dataset is presented in the classification report below (classes: 0 = TYLCV, 1 = Early Blight, 2 = Healthy).

UNDER PEER REVIEW

Table 2: Classification report

Class	Precision	Recall	F1-score	Support
0 (TYLCV)	0.94	0.86	0.90	51
1 (Early Blight)	0.92	0.96	0.94	51
2 (Healthy)	0.94	0.98	0.96	52
Accuracy			**0.94**	154
Macro Average	0.94	0.93	0.93	154

B. Image Processing for Inference:-

Before being fed into the model for inference, each input image undergoes a standardized preprocessing pipeline to ensure compatibility with the EfficientNetV2-B3 architecture and consistency with the training data distribution.The preprocessing steps include:

- **Resizing:**Each image is resized to match the fixed input dimensions required by EfficientNetV2-B3 (e.g., 300 x 300 pixels). This step is crucial because convolutional neural networks require a consistent input size across all images.
- **Conversion to Tensor:**Images are converted to numerical arrays of type float32, suitable for tensor operations within the deep learning framework.
- **Normalization:**Pixel values are normalized using the preprocess_input function provided by the EfficientNetV2 library. This adjusts the pixel distribution to match the ImageNet pre-training dataset, which stabilizes model activations and ensures reliable predictions.

This process guarantees that the images presented to the model during inference follow the same preprocessing conventions used during both fine-tuning and the original pre-training, ensuring accurate and consistent classification results.

C - Inference:-

Once the images have been preprocessed, they are passed to the fine-tuned model for classification. The model, loaded from its saved weights, processes each image as a normalized tensor compatible with the input requirements of the convolutional neural network architecture.

During inference, the model automatically extracts hierarchical visual features such as textures, color variations, and lesion patterns associated with different tomato leaf disease categories. Based on these learned internal representations, the network outputs a probability vector over all predefined classes in the dataset. Each value in this vector reflects the model's confidence for a given class. The final predicted label is obtained by selecting the class with the highest probability, following the standard argmax decision rule. This allows each input image to be automatically assigned to the most likely category among those known by the model.

Finally, the predicted probabilities can also be leveraged in the explainability stage to analyze which regions of the input image most strongly influenced the model's final decision, enabling a better understanding of class-specific discriminative patterns and supporting a more reliable interpretation of the model's behavior. Figure 2 illustrates the general inference workflow: a preprocessed image is fed into the model, which outputs a probability vector, and the predicted class is then selected.

UNDER PEER REVIEW IN IJAR

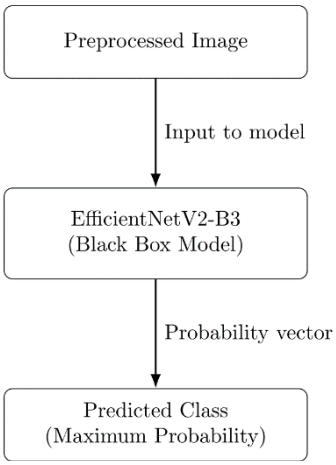


Figure 2:Inference workflow

D. XAI Method:-

In this work, explainable artificial intelligence (XAI) techniques are employed to interpret the predictions of the deep learning model. These methods aim to provide insights into the model's decision-making process by highlighting the most influential regions or features in the input images. Two main families of XAI methods are commonly distinguished in the literature**[24] [25]**:

- Global methodsprovide explanations at the level of the entire model, independently of any specific instance.
- Local methodsgenerate explanations specific to each individual input image.

In this study, two complementary local methods are selected: **Grad-CAM**, a gradient-based approach operating at the level of convolutional layers, and **LIME**, a perturbation-based method relying on a local surrogate model. Their selection is motivated by their theoretical complementarity, as Grad-CAM leverages model gradients while LIME relies on input perturbations, as well as their widespread adoption in the XAI community for computer vision applications. The combination of these techniques enables a more comprehensive understanding of the model behavior by providing both spatial and local explanations, thereby improving the transparency and reliability of the classification results.

GRAD-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) is a post-hoc explainability method introduced by Selvaraju et al. (2019)**[8]**. It leverages the gradients of a target class score with respect to the activations of a convolutional layer to generate a coarse localization map highlighting the most discriminative regions in the input image. The underlying intuition is that the spatial information preserved in convolutional feature maps, combined with gradient-based importance weighting, enables the identification of image regions that contribute the most to a specific prediction.

Formally, for a given target class c, the importance of each feature map $A^k$from a selected convolutional layer is computed by performing a global average pooling over the gradients of the class score y^c(prior to the softmax activation) with respect to the feature map activations:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

where Z denotes the total number of spatial locations in the feature map. The coefficientα_k^c thus captures the contribution of the k-th feature map to the prediction of class c.

Subsequently, the final class-discriminative localization map is obtained by computing a weighted linear combination of the feature maps, followed by the application of a rectified linear unit (ReLU):

UNDER PEER REVIEW IN IJAR

$$L^{c}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k} \alpha^{c}_{k} A^{k}\right)$$

The use of ReLU ensures that only the features that have a positive influence on the class of interest are retained, thereby improving the explainability of the resulting heatmap.

In practice, the last convolutional layer is typically selected for Grad-CAM, as it provides a good trade-off between high-level semantic information and spatial resolution. However, this choice is not universal. In certain scenarios, intermediate convolutional layers may yield more precise and informative explanations, particularly when the final layer produces overly coarse or semantically diffuse activations**[8] [9].**

In the context of this study, the final convolutional layer (top_conv) was initially evaluated as the default target for Grad-CAM visualization. However, preliminary experiments revealed that the resulting localization maps exhibited suboptimal alignment with the expected regions of interest. This limitation is primarily attributed to the high level of semantic abstraction at this depth of the network, which is typically associated with reduced spatial granularity and limited fine-grained localization capability.

To identify a more suitable target layer, a structured selection protocol was adopted based on three complementary criteria:

- Early-stage convolutional layers were due to their limited semantic representation capacity, as they predominantly encode low-level visual primitives such as edges, textures, and local contrast patterns**[26]**.
- Normalization and non-spatial transformation layers were excluded because they do not preserve meaningful spatial activation maps required for gradient-based localization methods such as Grad-CAM.
- The selection process was constrained to the project_conv layers of the network, with particular emphasis on the final MBConv stage (block6), which provides a favorable trade-off between semantic richness and spatial resolution**[27]**.

Empirical analysis revealed that intermediate layers within this stage, particularly block6g_project_conv, produced localization maps that were both spatially coherent and semantically consistent with the discriminative regions relevant to the classification task.

Consequently, block6g_project_conv was selected as the target layer for subsequent Grad-CAM analyses. This observation highlights the critical importance of intermediate representation selection in convolutional architectures, particularly when applying gradient-based explainability techniques to fine-grained visual recognition tasks.

LIME

LIME (Local Interpretable Model-agnostic Explanations), introduced by Ribeiro et al. (2016)**[10]**, is a post-hoc local explainability method designed to provide interpretable approximations of complex machine learning models. Being model-agnostic, LIME can be applied to any predictive model without requiring access to its internal structure or parameters.

The core idea of LIME is to approximate the behavior of a black-box model f in the vicinity of a given instance x by learning an interpretable surrogate model $g \in G$, typically chosen from a family of simple models such as sparse linear regressors. This approximation is obtained by solving the following optimization problem:

$$\xi(x) = \arg\min_{g \in G} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

where $\mathcal{L}(f, g, \pi_x)$is a locality-aware loss function that measures how well the surrogate model g approximates the predictions of f in the neighborhood of x, weighted by a proximity kernel π_x. The term $\Omega(g)$ denotes a complexity penalty that enforces interpretability by favoring sparse and simple models.

In the context of image data, the interpretable representation is typically defined in terms of superpixels, which partition the image into semantically meaningful regions. The surrogate model g thus operates on a binary representation indicating the presence or absence of each superpixel, allowing the attribution of importance weights to different regions of the image.

In this study, LIME is implemented using a LASSO (Least Absolute Shrinkage and Selection Operator) linear surrogate model. Explanations are generated using N = 1000 perturbed samples and S = 50 superpixels. These parameter choices are consistent with commonly adopted configurations in image-based XAI applications, including plant disease classification tasks, as they provide a practical trade-off between explanation stability and computational efficiency**[11].**

UNDER PEER REVIEW IN IJAR

E. Qualitative Evaluation:-

Qualitative evaluation consists of analyzing a model's outputs through human or visual inspection, with a focus on explainability, coherence, and relevance, rather than relying on numerical metrics. In the context of explainable artificial intelligence (XAI), it aims to assess whether the generated explanations provide sufficient information to understand the classifier's decision-making process and whether the highlighted elements are consistent with domain knowledge.

It addresses two key questions: (i) Access the Classifier, which evaluates the ability to understand the model's logic, and (ii) Find Alignment, which measures the alignment between model explanations and human reasoning**[12]**.

1. Visual explanations

The figures presented below are extracted from the qualitative evaluation conducted on the test set. They illustrate representative samples selected during the inference phase in order to assess the visual explainability of the proposed approach.

Table 3: Visual explanation of Model Decisions Using LIME Masks

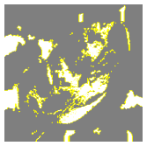
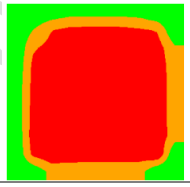
Class	Input	GRAD CAM	LIME
Healthy			
Early blight			
TYLCV			

Table 4: Visual explanation of Model Decisions using GRAD-CAM Heatmap

For each class, one representative image is selected from the test set in order to qualitatively assess the explainability of the model predictions. The selected samples are then analyzed using two post-hoc explainability methods: LIME and Grad-CAM. For each image, we provide:

- the original input image,
- the LIME mask highlighting influential superpixels,
- the Grad-CAM heatmap indicating class-discriminative activation regions.

These visualizations are presented in a comparative framework to assess the consistency and complementarity of the explanations**[13].**

2. Analysis

The generated visualizations highlight complementary explanatory behaviors between the two investigated approaches. LIME masks isolate locally discriminative regions in the form of superpixels, whereas Grad-CAM produces spatially smoother activation maps that are more consistent with the deep internal representations of the convolutional network.

UNDER PEER REVIEW IJAR

This difference results in two distinct levels of interpretation: a fine-grained and localized analysis of contributive regions with LIME, and a more global representation of semantically relevant areas with Grad-CAM. The combined use of both approaches therefore provides a more robust qualitative explainability of the model behavior by jointly capturing attributional precision and spatial coherence of activations.

F. Quantitative Evaluation:-

The evaluation of explanation maps produced by XAI methods such as LIME and Grad-CAM is a crucial step for assessing the quality of explanations and their actual influence on model predictions.

The literature proposes several quantitative metrics for this purpose. In this study, we focus on two main properties: **fidelity**, which evaluates to what extent the highlighted regions effectively contribute to the model's prediction, and **stability**, which measures the robustness of explanations under small input perturbations.

Fidelity is assessed jointly using IAUC and DAUC metrics, which are based on progressive insertion and deletion of the most important regions identified in the explanation maps.

- IAUC (Insertion Area Under Curve): this metric measures the progressive increase in model confidence as the most important regions of the explanation map are gradually reintroduced into the input image. It evaluates the ability of the explanation method to correctly identify the most discriminative regions contributing to the prediction**[15].**
- DAUC (Deletion Area Under Curve): this metric measures the progressive decrease in model confidence as the most important regions are gradually removed from the input image. A strong drop in confidence indicates that the identified regions are highly influential in the model's decision**[15].**
- Stability: this metric evaluates the robustness of explanations under small input perturbations. High stability indicates that minor variations in the input image do not significantly affect the generated explanation maps, reflecting consistent and reproducible explanations**[17]**.

Although other metrics exist, this selection is motivated by their relevance in image classification tasks as well as their widespread use in previous works**[14] [15] [16][17].**

1. Fidelity & Stability

The objective of this section is to complement the qualitative analysis presented previously with a quantitative evaluation of the explainability methods. This analysis aims to assess the importance of the selected regions on the model's decision as well as the stability of the prediction under controlled perturbations.

The evaluation is conducted on a stratified subset corresponding to 20% of the test samples from each class. The subset is randomly selected using a fixed random seed to ensure reproducibility, while maintaining an equal class distribution in order to avoid class imbalance bias during the evaluation. This experimental design provides a trade-off between computational cost and statistical representativeness.

Table 5: Quantitative evaluation of explanation methods using DAUC, IAUC, and Stability metrics

Method	DAUC↓	IAUC↑	Stability↓
GRAD CAM	0.5134	0.8921	**0.0002**
LIME	**0.4187**	**0.9268**	0.0025

2. Analysis

The quantitative results reported in Table 3 highlight differences between Grad-CAM and LIME in terms of both fidelity and stability of the generated explanations.

First, LIME consistently outperforms Grad-CAM on the perturbation-based metrics. It achieves a lower Deletion AUC (0.4187 vs. 0.5134), indicating that the regions identified by LIME are more critical to the model's decision, as their removal leads to a stronger degradation in prediction confidence.It should be noted that the absolute DAUC values remain relatively high. However, as highlighted by Tristan Gomez et al. (2022) **[16]**, the relevance of these metrics lies primarily in the relative comparison between methods rather than in their absolute values.

Complementarily, LIME obtains a higher Insertion AUC (0.9268 vs. 0.8921), reflecting a better ability to reconstruct the model's prediction from the most relevant regions.

Regarding stability, the employed metric corresponds to the ΔL_2 distance between explanation maps under noisy perturbations. In this context, lower values indicate higher stability. Grad-CAM therefore exhibits better stability (0.0002 vs. 0.0025 for LIME), suggesting that its attribution maps are less sensitive to input variations.

UNDER PEER REVIEW IN IJAR

Overall, these results highlight a clear trade-off between fidelity and stability: LIME produces explanations that are more faithful to the model's decision process, while Grad-CAM generates more stable explanations under perturbations.

G. Discussion:-

1- Quantitative and Qualitative Evaluation and Positioning in the Literature

In our study, LIME achieves a higher IAUC than Grad-CAM across the majority of tested disease classes, which quantitatively confirms its ability to isolate regions that are causally relevant rather than merely correlated with the prediction. These results are consistent with those reported by Petsiuk et al. (2018)**[15]**, who demonstrate that perturbation-based methods tend to produce attributions that are more faithful with respect to such metrics.

Furthermore, qualitative observations reveal a correlation between visually relevant regions (leaf lesions) and the discriminative regions identified by both methods. This finding aligns with studies showing that robust models tend to better align with expert human perception**[14]**. Although explainability remains partly subjective, research from cognitive science and decision psychology suggests that certain properties of explanations can be evaluated objectively**[18]**. The quantitative results obtained in this study further support this hypothesis.

2- Stability and fidelity of explanations

The stability of explainability methods is a central issue in XAI, as the instability of interpretations directly affects the trust placed in deep learning systems**[14]**. The results obtained in this study reveal an intrinsic tension between fidelity and stability of explanations, a trade-off reported by Yeh et al. (2019)**[21]** and empirically observed in our experiments within the context of EfficientNet applied to agricultural crops (tomato leaves).

Grad-CAM produces spatially coherent activation maps, whose stability can be explained by the deterministic nature of backpropagation through the deep convolutional layers of EfficientNet: MBConv blocks, by aggregating information through depthwise separable convolutions**[19]**, tend to concentrate activations on semantically stable regions. In contrast, LIME segments the image into superpixels and samples a local perturbation space to fit a linearly interpretable model**[10]**; this stochastic procedure provides higher decision granularity, identifying more precisely the regions actually required for the prediction, at the cost of higher variability across runs. This result is consistent with the findings of Alvarez-Melis & Jaakkola (2018)**[20]**, who show that perturbation-based methods are structurally prone to instability due to the high dimensionality of the sampling space.

The combination of Grad-CAM and LIME thus contributes to a complementary interpretability framework, where Grad-CAM provides a global and stable view of discriminative regions, while LIME refines the analysis by isolating the superpixels causally involved in the decision. This complementarity does not merely reflect a methodological choice; it reveals the inherent duality of any explanation between representational coherence and attributional precision.

3- Failure Cases

Although both Grad-CAM and LIME generally produce explanations consistent with visible disease symptoms, several failure cases were identified:

- Inaccurate localization of disease-related regions;
- Partial coverage of visible symptoms;
- Disagreement between Grad-CAM and LIME explanations.

These observations highlight the limitations of post-hoc explainability techniques and reinforce the importance of complementing qualitative inspection with quantitative evaluation metrics.

4- Limitations and future work

A major limitation of this study is the absence of evaluation involving domain experts in agriculture. The DAUC and IAUC metrics assess the internal consistency of explanations with respect to the model, but do not capture their interpretability for end users. A human evaluation protocol, for instance a Which-Is-Better (WIB) survey comparing Grad-CAM and LIME explanations provided to agronomists, would allow measuring the alignment between computational saliency and perceived relevance. This aspect is essential to validate the deployment of such systems in real-world conditions.

The issue of Out-of-Distribution (OOD) data remains an open challenge. AUC-based metrics rely on the assumption that the model evaluates images drawn from the same distribution as the training set; however, perturbed images generated during deletion and insertion procedures may fall outside this distribution, introducing bias in confidence estimation. This phenomenon, documented by Hooker et al. (2019)**[23]** under the ROAR (RemOveAnd Retrain)

UNDER PEER REVIEW IN IJAR

framework, represents a methodological limitation. The integration of complementary metrics such as Deletion Correlation (DC) and Insertion Correlation (IC) could provide a more reliable assessment of external validity.
An additional limitation concerns the generalizability of the conclusions to other architectures or agricultural datasets. The observed trade-off between fidelity and stability is specific to EfficientNetV2 on the considered dataset; a systematic transferability study, for instance on ResNet or Vision Transformers (ViT), would be required to determine whether this trade-off is a general property or dependent on the inductive bias of the architecture.

Conclusion:-

This work is part of the broader challenge of explainability in deep learning models applied to tomato leaf disease classification, a domain in which end-user trust directly conditions the adoption of artificial intelligence systems. Given the inherent opacity of deep neural networks, post-hoc explainability has become an essential approach for reconciling predictive performance with decision intelligibility.
This study investigated the complementarity of Grad-CAM and LIME for interpreting an EfficientNetV2B3 model applied to plant disease classification. Through a combined qualitative and quantitative evaluation, the study highlighted an empirical trade-off between stability and fidelity of explanations, and demonstrated that the identified discriminative regions effectively correspond to symptoms characteristic of the studied pathologies, a necessary condition for expert user trust.
Building on these findings, integrating explainability constraints directly into the training loop, as proposed by Ross et al. (2017)**[22]**, represents a promising direction to better align learned representations with expert domain knowledge moving explainability from a post-hoc analysis tool to an intrinsic property of model design.

References:-

[1] RibanaRoscher, Bastian Bohn, Marco F. Duarte, Jochen Garcke (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8, 42200–42216. https://doi.org/10.48550/arXiv.1905.08883
[2] Jordan D. Fuhrman, Naveena Gorre, Qiyuan Hu, Hui Li, Issam El Naqa, Maryellen L. Giger (2022). A review of explainable and interpretable AI with applications in COVID-19 imaging. *Medical Physics*, 49(1), 1–14. https://dx.doi.org/10.1002/mp.15359
[3] Melkamu Mersha, Khang Lam, Joseph Wood, Ali AlShami, Jugal Kalita (2025). Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *arXiv*. https://dx.doi.org/10.48550/arXiv.2409.00265
[4] Nguyen Van Tu, Pham Nguyen Hai Long, Vo Hoai Viet (2025). xAI-CV: An overview of explainable artificial intelligence in computer vision. *arXiv*, arXiv:2509.18913v1 [cs.CV]. https://dx.doi.org/10.48550/arXiv.2509.18913
[5] Apoorva Aravindkumar, MarimuthuRamadoss, Saqhibuddeen Ahmed Fakhruddin Ahmed, Vidhya Sampath, Kishor Lakshminarayanan (2026). Explainable AI in healthcare: A systematic review of XAI use cases in imaging, diagnostics, and rehabilitation. *Frontiers in Artificial Intelligence*, 9.https://dx.doi.org/10.3389/frai.2026.1749527
[6] Jyothi Peta, Srinivas Koppu (2024). Explainable Soft Attentive EfficientNet for breast cancer classification in histopathological images. *Biomedical Signal Processing and Control*, 90, 105828. https://dx.doi.org/10.1016/j.bspc.2023.105828
[7] Kaihua Wei, Bojian Chen, Jingcheng Zhang, Shanhui Fan, Kaihua Wu, Guangyu Liu, Dongmei Chen (2022). Explainable deep learning study for leaf disease classification. *Agronomy*, 12(5), 1035. https://dx.doi.org/10.3390/agronomy12051035
[8] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra (2019). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *arXiv*, arXiv:1610.02391v4 [cs.CV]. https://dx.doi.org/10.48550/arXiv.1610.02391
[9] MykytaSkliarov, Radwa El Shawi, ChediaDhaoui, Nada Ahmed (2025). A comparative evaluation of explainability techniques for image data. *Scientific Reports*, 41898 (2025). https://www.nature.com/articles/s41598-025-25839-y
[10] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv preprint*, arXiv:1602.04938v3 [cs.LG]. https://doi.org/10.48550/arXiv.1602.04938
[11] Trong-Minh Hoang, Van-Hau Bui, Van-Son Nguyen, Duc-Thang Doan, Hoang-Anh Dang, Anh-Thu Pham (2026). A comprehensive evaluation of lightweight deep learning models for tomato disease classification on edge computing environments. *Scientific Reports*,12320 (2026). https://doi.org/10.1038/s41598-026-42439-6
[12] NourahAlangari, Mohamed El Bachir Menai, Has-sanMathkour, Ibrahim Almosallam (2023). Exploring

UNDER PEER REVIEW IN IJAR

Evaluation Methods for Interpretable Machine Learn-ing: *A Survey*. *Information*,2023, 14(8), 469. https://doi.org/10.3390/info14080469
[13] Sadia Nazim, Muhammad Mansoor Alam, SyedSafdar Rizvi, Jawahir Che Mustapha, Syed ShujaaHussain, MazlihamMohd Suud (2025). Advancingmalware imagery classication with explainable deeplearning: A state-of-the-art approach using SHAP,LIME and Grad-CAM. *PLOS One*. https://doi.org/10.1371/journal.pone.0318542
[14] NourahAlangari, Mohamed El Bachir Menai, Has-sanMathkour, Ibrahim Almosallam (2023). Exploring Evaluation Methods for Interpretable Machine Learn-ing: *A Survey*. Information,2023, 14(8), 469. https://doi.org/10.3390/info14080469
[15] Vitali Petsiuk, Abir Das, Kate Saenko (2018).RISE:Randomized Input Sampling for Explanation of Black-box Models. *arXiv*, arXiv:1806.07421v3 [cs.CV].https://doi.org/10.48550/arXiv.1806.07421
[16] Tristan Gomez, Thomas FrØour, HaroldMouchŁre (2022).Metrics for Saliency MapEvaluation of Deep Learning ExplanationMethods. *Conference paper*, pp 8495.https://doi.org/10.1007/978-3-031-09037-0_8
[17] Chirag Agarwal, Nari Johnson, Martin Pawel-czyk, Satyapriya Krishna, EshikaSaxena,MarinkaZitnik, HimabinduLakkaraju (2022).Re-thinking Stability for Attribution-based Expla-nations. *arXiv*, arXiv:2203.06877v1 [cs.LG].https://doi.org/10.48550/arXiv.2203.06877
[18] An-phi Nguyen, María Rodríguez Martínez (2020).On quantitative aspects of model inter-pretability. *arXiv*, arXiv:2007.07584v1 [cs.LG]. https://doi.org/10.48550/arXiv.2007.07584
[19] Mingxing Tan, Quoc V. Le (2019).E-cientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv*, arXiv:1905.11946v5 [cs.LG]. https://doi.org/10.48550/arXiv.1905.11946
[20] David Alvarez-Melis, Tommi S. Jaakkola (2018).Towards Robust Interpretability with Self-ExplainingNeural Networks. *arXiv*, arXiv:1806.07538v2 [cs.LG].https://doi.org/10.48550/arXiv.1806.07538
[21] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, Pradeep Ravikumar (2019).On the (In)delity and Sensitivity for Explanations.*arXiv*, arXiv:1901.09392v4 [cs.LG]. https://doi.org/10.48550/arXiv.1901.09392
[22] Andrew Slavin Ross, Michael C. Hughes, Finale Doshi-Velez (2017).Right for the Right Rea-sons: Training Dierentiable Models by Constraining their Explanations. *arXiv*, arXiv:1901.09392v4[cs.LG]. https://doi.org/10.48550/arXiv.1703.03717
[23] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, Been Kim (2018).A Benchmark for Interpretability Methods in Deep Neural Networks.*arXiv*, arXiv:1806.10758v3 [cs.LG]. https://doi.org/10.48550/arXiv.1806.10758
[24] Rabia Saleem, Bo Yuan, Fatih Kurugollu, AshiqAnjum, Lu Liu (2022).Explaining deep neural networks: A survey on the global interpretation methods.*Neurocomputing*, Volume 513, 7 November 2022,Pages 165-180. https://doi.org/10.1016/j.neucom.2022.09.129
[25] Gesina Schwalbe, Bettina Finzel (2023).A comprehensive taxonomy for explainable artificialintelligence: a systematic survey of surveys on methods and concepts. Data Mining andKnowledgeDiscovery, Volume38, pages 3043-3101, (2024). https://doi.org/10.1007/s10618-022-00867-8
[26] Matthew D. Zeiler, Rob Fergus (2014).Visualizingand Understanding Convolutional Networks. *Conference paper*, pp 818-833.
[27] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra (2016).Grad-CAM: Visual Explanationsfrom Deep Networks via Gradient-based Localization.*arXiv*, arXiv:1610.02391v4 [cs.CV]. https://doi.org/10.48550/arXiv.1610.02391
[28] Donatien Gueswendé Kabore, Cheikh Sarr and Amadou MbagnickGning (2026). Efficient and accurate detection of tomato leaf diseases using fine-tuned EfficientNetV2 in Keras,1819-6608, 1785-1791, Volume No.20. Issue No.20https://doi.org/10.59018/1025201

UNDER PEER REVIEW IN IJAR