Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume *, Number *, Spring 20**, pp. **-**
Received: **/**/**, Revised: **/**/**, Accepted: **/**/**

Towards Trustworthy Diagnostics: An Interpretable Framework for Uncovering Shortcut Learning in Plant Disease Models

Babak Masoudi[1*], Safoora Bazzi[2]
1*- Department of information technology, Payame Noor University (PNU), P.O.Box 19395-3697, Tehran, I.R of Iran.
2- Department of Biology, Payame Noor University (PNU), P.O.Box 19395-3697 Tehran, I.R of Iran.
[1*]b.masoudi@pnu.ac.ir, [2]s.bazzi@pnu.ac.ir

Corresponding author's address: Babak Masoudi, Department of information technology, Payame Noor University (PNU), P.O.Box 19395-3697, Tehran, I.R of Iran.

Abstract- The deployment of artificial intelligence in agriculture requires models that are not only accurate but also transparent and trustworthy. While deep learning models have demonstrated impressive performance on benchmark datasets, their internal validity and generalization capabilities are often assumed rather than verified. This study provides a methodological investigation into the gap between high accuracy and genuine diagnostic reliability, particularly within the context of standardized, laboratory-style datasets. Our investigation began with the development of a novel hybrid architecture that achieved a high accuracy of 98.61% on the PlantVillage dataset. However, a subsequent interpretability analysis using Layer-GradCAM revealed that the model's success was substantially reliant on spurious background features, a behavior commonly known as shortcut learning. To systematically investigate this phenomenon, we conducted further experiments with strongly regularized models, including a pure CNN baseline. Although these models maintained high accuracy and even improved performance on diagnostically challenging classes, the shortcut learning phenomenon, while mitigated, was not entirely eliminated. Our findings indicate that high accuracy on benchmark datasets like PlantVillage can be a deceptive metric for assessing real-world trustworthiness. This work reframes the challenge in this domain, arguing that rigorous, XAI-driven validation should be an integral part of model evaluation, taking precedence over the pursuit of incremental accuracy gains on potentially biased datasets. Ultimately**,** we contribute a critical analysis into the limitations of such benchmarks and underscore the essential role of interpretability in developing reliable AI for agricultural applications.

Keywords- Explainable Artificial Intelligence (XAI), Shortcut Learning, Trustworthy AI, Deep Learning, Computational Plant Science, Model Interpretability.

I. INTRODUCTION

Global food security is intrinsically linked to agricultural productivity, a sector constantly under threat from plant diseases. Pathogens such as fungi, bacteria, and viruses can lead to devastating crop losses. However, conventional disease identification often relies on manual visual inspection, a process with notable limitations. It can be subjective, labor-intensive, and difficult to scale across large agricultural areas. Furthermore, accurately differentiating between diseases with similar visual symptoms requires extensive expertise, making it a challenging task for individuals without specialized training [1].

The rapid advancements in artificial intelligence (AI), particularly in the domain of deep learning (DL), have paved the way for automated systems for plant disease diagnostics. The field has evolved from earlier machine learning approaches to modern deep convolutional neural networks (CNNs) that enable end-to-end automated feature extraction and classification [2], [3]. Numerous studies have demonstrated the promise of DL, reporting high classification accuracies on the popular PlantVillage dataset under controlled laboratory conditions [4], [5].

However, the pursuit of incrementally higher accuracy

scores often overshadows a critical question: does a model's performance rely on learning pathological features, or does it exploit statistical artifacts within benchmark datasets? High performance on controlled, laboratory-style datasets does not guarantee generalization to real-world agricultural environments. More alarmingly, a model might achieve exceptional accuracy by learning spurious correlations, for instance, associating a specific background texture or lighting artifact with a particular disease class. This phenomenon, known as shortcut learning, can lead to models that appear successful in testing but are brittle and unreliable when deployed in the field [6]. The risks associated with such behavior have been compellingly demonstrated in high-stakes medical domains, where models have been found to rely on imaging artifacts rather than true pathology, highlighting the need for similar critical investigations in agriculture [7], [8]

The inherent "black box" nature of these powerful models represents a significant barrier to their adoption in high-stakes domains such as agriculture. For a diagnosis to be trusted by an agronomist or farmer, providing a classification label alone is insufficient. It is essential to build systems that are transparent in their reasoning. This necessity has spurred the growth of Explainable Artificial Intelligence (XAI), a field dedicated to developing methods that render the decisions of complex models understandable to humans [9]. Notably, XAI techniques provide the tools to "unmask" these shortcut strategies and validate what a model has actually learned [10].

This study provides a methodological investigation into the problem of shortcut learning in plant disease classification, using the widely-cited PlantVillage dataset as a representative case study. Our research began with the standard goal of developing a high-performance, interpretable model. However, our interpretability analysis revealed that the model's impressive accuracy was deceptive, relying heavily on non-pathological background features. This discovery fundamentally shifted the focus of our research. Instead of merely proposing another high-accuracy model, this paper provides a critical case study on the discordance between accuracy and trustworthiness.

Therefore, the primary contributions of this study are reframed as follows:

1. A systematic demonstration that deep learning models trained on the PlantVillage benchmark are highly susceptible to shortcut learning by relying on spurious background features.

2. A comparative analysis showing this behavior persists across different architectures and training strategies, suggesting the issue is deeply embedded within the dataset.

3. An evidence-based argument that for high-stakes applications like agriculture, XAI-driven validation must be an integral component of the model evaluation pipeline.

II. Related Work

This section reviews the existing literature from three crucial perspectives that form the foundation of our research: (1) the pervasive challenge of shortcut learning in deep learning, (2) the role of Explainable AI (XAI) as a tool for model validation rather than just visualization, and (3) the specific critiques and known limitations of benchmark datasets in agriculture, particularly PlantVillage.

A. Shortcut Learning: A Fundamental Challenge in AI

Shortcut learning, sometimes referred to as the "Clever Hans" effect, describes the phenomenon where deep learning models learn to exploit unintended and overly simple cues in the training data instead of the underlying, generalizable features [6] . These models often achieve high performance on standard benchmarks but fail unpredictably when deployed in real-world scenarios with slightly different data distributions. This challenge is not confined to a single domain; it is a fundamental issue across artificial intelligence. For instance, recent studies have highlighted this problem in advanced Large Language Models (LLMs) used for safety-critical applications [11].

The most compelling evidence for the risks of shortcut learning comes from high-stakes medical imaging applications. In a landmark study, Nauta et al. [7] demonstrated that a skin cancer classifier learned to associate the presence of benign lesions with colored calibration patches that only appeared in images of benign cases. Similarly, Wallis and Buvat [8] uncovered a bias in a widely-used brain tumor dataset where implicit radiologist choices allowed a model to achieve high accuracy without analyzing the tumor itself. These studies serve as a stark warning: without critical validation, high accuracy can be a dangerously misleading metric. While the risks in medicine are well-documented, a similar systematic investigation into the prevalence and nature of shortcut learning in agricultural benchmarks remains a less-explored but equally critical area.

B. Explainable AI (XAI) as a Diagnostic Tool

The "black box" nature of deep learning has spurred the development of Explainable AI (XAI), a field focused on making model decisions transparent and interpretable [9]. In the context of agriculture, there is a growing trend of incorporating XAI techniques to increase trust and validate model outputs. Several recent studies have employed methods like Grad-CAM, LIME, or SHAP to visualize which parts of an image a model focuses on when making a diagnosis [12-14].

However, as a recent survey on XAI in plant health monitoring suggests, the aspect of explainability is often neglected or applied superficially [15]. The predominant use of XAI in the current agricultural literature is confirmatory; it is used to show that a high-performing model is focusing on the "correct" regions, thereby confirming its success. This approach, while useful, often

misses the more powerful application of XAI as a diagnostic tool for debugging. As argued by Lapuschkin et al. [10], the true value of these techniques lies in their ability to "unmask" flawed reasoning and discover what a model has *really* learned. The use of XAI in a critical, investigative capacity to actively uncover shortcut strategies in high-accuracy agricultural models represents a significant research gap that our study aims to address.

C. *Critiques of Standardized Agricultural Benchmarks*

The performance of any deep learning model is intrinsically linked to the data it is trained on. While benchmark datasets like PlantVillage have been instrumental in advancing research [4], [5], their limitations are becoming increasingly apparent. A key issue with many visual datasets is the presence of inherent biases, which can stem from the data collection and curation process [16].

Specifically for datasets created in controlled, laboratory-style settings, the lack of background variation, consistent lighting, and uniform leaf positioning can introduce statistical artifacts. These artifacts provide "easy" shortcuts for models to learn, which do not exist in real-world farm conditions. This issue touches upon a deeper question about the values embedded in dataset creation, as explored by Scheuerman et al. [17], who argue that the pursuit of efficiency and universality in dataset development can come at the expense of context and careful representation. While the general limitations of PlantVillage are often acknowledged, a systematic study that provides direct, visual evidence linking the dataset's characteristics to specific, flawed model behaviors (i.e., shortcut learning) is lacking. Our work directly bridges this gap.

In summary, the existing literature establishes shortcut learning as a major threat, highlights the potential of XAI as a diagnostic tool, and points to the limitations of standard benchmarks. This study synthesizes these three areas by conducting a methodological investigation that uses XAI to systematically uncover and analyze shortcut learning within the context of the PlantVillage dataset.

III. Material and Methods

The methodological framework of this study is designed as a three-stage investigation to progressively uncover and analyze the phenomenon of shortcut learning in plant disease classification. The overall workflow is illustrated in Fig. 1. We first established a baseline performance with a novel hybrid architecture that served as our primary investigative tool. Following a critical analysis of its interpretability, we then conducted two subsequent experiments involving strong regularization and a simplified baseline to systematically diagnose the root cause of the observed issues.

A. *Dataset and Preprocessing*

This study utilized the widely-used PlantVillage dataset, an open-access image repository introduced in [18]. To ensure computational efficiency and strict reproducibility, we created a fixed 15% subset via stratified random sampling while preserving the original class distribution. This subset contains 8,129 color images, encompassing 38 distinct classes that represent 14 plant species afflicted by various diseases, alongside their healthy counterparts.

The subset was partitioned into three distinct sets using a 70%/15%/15% split ratio with stratification:

- Training Set: 5,690 images.
- Validation Set: 1,219 images.
- Test Set: 1,220 images, held out exclusively for final, unbiased evaluation.

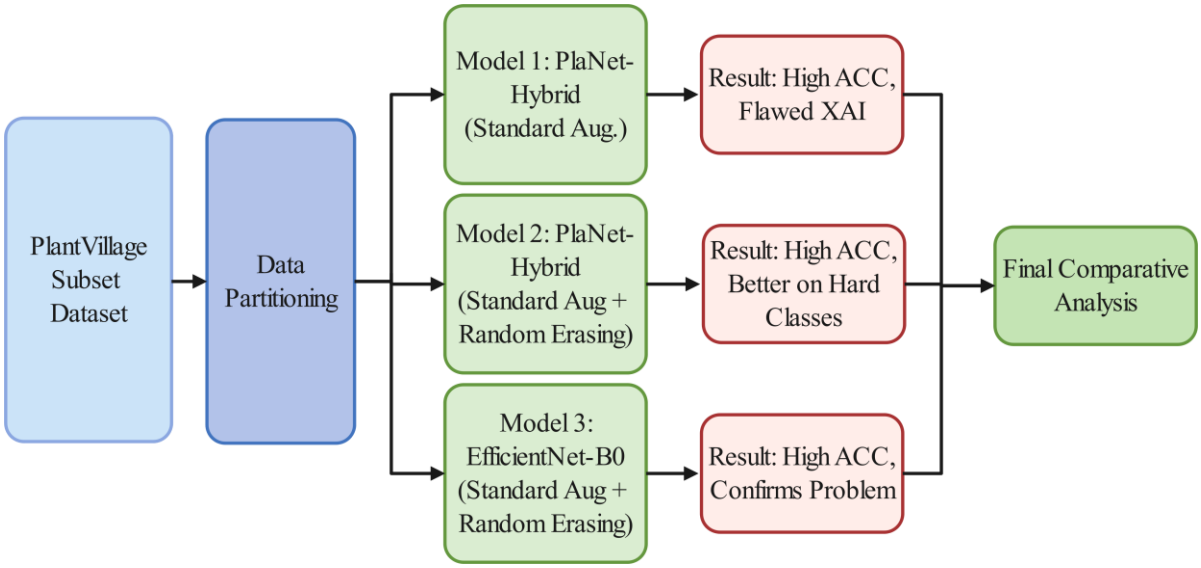

Fig. 1. The workflow of this study.

Two distinct data augmentation pipelines were employed. For the initial baseline model (Experiment 1), a standard augmentation strategy was used (e.g., random resized cropping, horizontal flipping, and color jittering). For the subsequent investigative experiments (Experiments 2 and 3), we implemented a much stronger regularization pipeline designed to combat shortcut learning. This pipeline was enhanced with the addition of RandomErasing [19]. This technique randomly occludes a region of the input image, preventing the model from over-relying on any single, localized feature (such as a background artifact) and forcing it to learn a more distributed representation. A visual comparison of these data transformation pipelines is presented in Fig. 2. All images were resized to 224×224 pixels and normalized using ImageNet statistics.

B. Model Configurations for Investigation

To systematically investigate the phenomenon of shortcut learning, this study evaluates three distinct model configurations.

1) PlaNet-Hybrid: An Investigative Hybrid Model

Our initial model is a novel dual-branch hybrid architecture designed to test the hypothesis that fusing local and global features could yield superior performance and potentially more robust representations. As depicted in Fig. 3, it consists of:

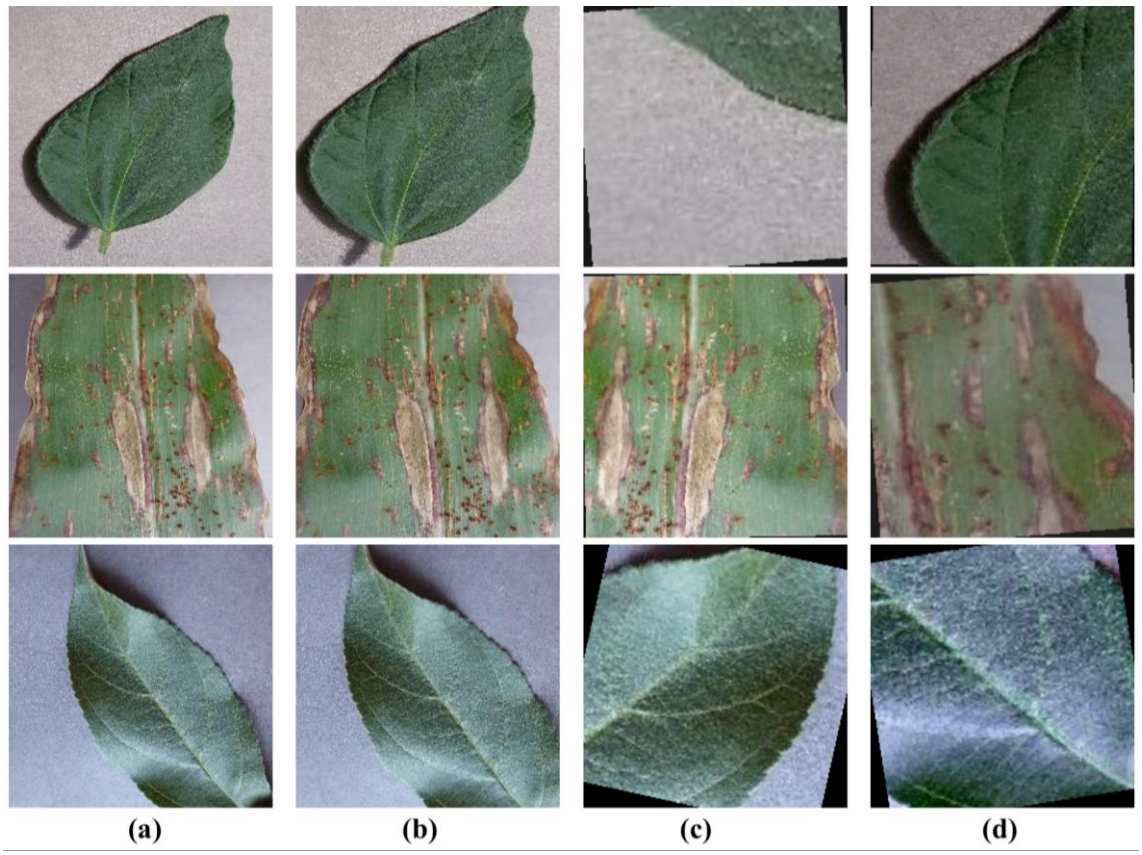

Fig. 2. Visual comparison of image preprocessing pipelines. (a) Original images. (b) Deterministic validation/test preprocessing (resize and center crop). (c) Standard training augmentation (used for Experiment 1). (d) Strong regularizing augmentation with RandomErasing .

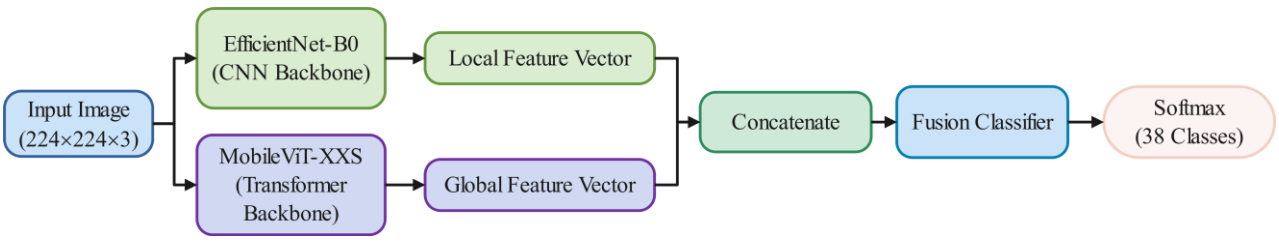

Fig. 3. Schematic of the proposed PlaNet-Hybrid architecture.

- A local feature branch: Utilizing a pretrained EfficientNet-B0 [20] to capture fine-grained patterns.
- A global context branch: Employing a pretrained MobileViT-XXS [21] to model broader spatial relationships.
- A fusion classifier: The feature vectors from both branches are concatenated and passed through a classification head with Batch Normalization and Dropout for regularization.

2) Regularized Model: planet_hybrid_regularized

This model uses the exact same PlaNet-Hybrid architecture as the baseline. The only difference lies in the training methodology, where it is trained using the strong augmentation pipeline featuring RandomErasing.

3) Pure CNN Baseline: efficientnet_b0_regularized

To isolate the effect of the architecture, this model consists of a standard, pretrained EfficientNet-B0. It was also trained using the strong augmentation pipeline, allowing for a direct comparison of architectural susceptibility to shortcut learning under identical training conditions.

C. Explainable AI (XAI) for Model Interpretation

To understand *how* our models were making decisions, we employed the gradient-based XAI technique LayerGradCam. This method generates a heatmap that highlights the image regions most influential to a model's prediction.

A systematic pre-analysis was conducted to select the optimal target layer for attribution. We found that for the EfficientNet-based architectures, the final convolutional layer in the 5th block (blocks[5]) provided the best trade-off between high-level semantic features and sufficient spatial resolution. Furthermore, to enhance clarity and focus, we post-processed the generated heatmaps. We applied a percentile threshold to the raw attribution map, retaining only the top 10-15% of the highest-intensity activations. We empirically confirmed that this threshold provides an excellent balance between removing low-level noise and preserving the core explanatory features, thereby pinpointing the model's most critical regions of focus with greater precision.

D. Experimental Setup and Evaluation Metrics

All models were trained within the PyTorch framework for a maximum of 25 epochs, utilizing an early stopping mechanism (patience=5). We employed the AdamW optimizer with a learning rate of 1e-3 and a Cosine Annealing scheduler. Model performance was quantified using Overall Accuracy, Precision, Recall, F1-Score (Macro and Weighted averages), and the Area Under the ROC Curve (AUC) on the held-out test set to ensure a fair and robust comparison.

IV. RESULTS

The results of our three-stage investigation are presented sequentially. We first report the high-accuracy performance of our baseline model. We then present the critical interpretability analysis that revealed the shortcut learning phenomenon, followed by the comparative results of our subsequent regularization experiments**.**

A. Experiment 1: High-Accuracy Performance of the Baseline Model

The initial experiment established a performance benchmark using our PlaNet-Hybrid architecture (Model 1) with a standard training approach. On the held-out test set, the model achieved a high overall accuracy of 98.61% and a macro-average F1-score of 97.98% (Table I). These metrics are comparable to those reported in existing literature for the PlantVillage dataset. On the surface, these results suggested that the model was highly effective.

B. Interpretability Analysis Reveals Flawed Reasoning

Despite the strong performance metrics, the XAI validation exposed a discrepancy between the model's accuracy and its diagnostic reasoning. As visualized for several correctly classified examples in Fig. 4, the Layer-GradCAM heatmaps for the baseline model (Model 1) frequently failed to focus on the pathological symptoms. Instead, its attention was often diffused or concentrated on spurious features in the image background. This provided strong evidence that the model was not performing genuine disease diagnosis but was exploiting dataset shortcuts, rendering its performance fundamentally untrustworthy.

TABLE I
COMPARATIVE PERFORMANCE METRICS (OVERALL ACCURACY AND MACRO/WEIGHTED F1-SCORES) FOR ALL THREE EXPERIMENTAL MODELS ON THE HELD-OUT TEST SET.

Model / Experiment	Architecture	Training Strategy	Test Accuracy	Macro Avg F1-Score	Weighted Avg F1-Score
Model 1	**PlaNet-Hybrid**	Standard Augmentation	**98.61%**	**97.98%**	**98.59%**
Model 2	**PlaNet-Hybrid**	Strong Regularization (+ *RandomErasing*)	**98.61%**	97.47%	**98.59%**
Model 3	EfficientNet-B0	Strong Regularization (+ *RandomErasing*)	98.28%	97.83%	98.27%

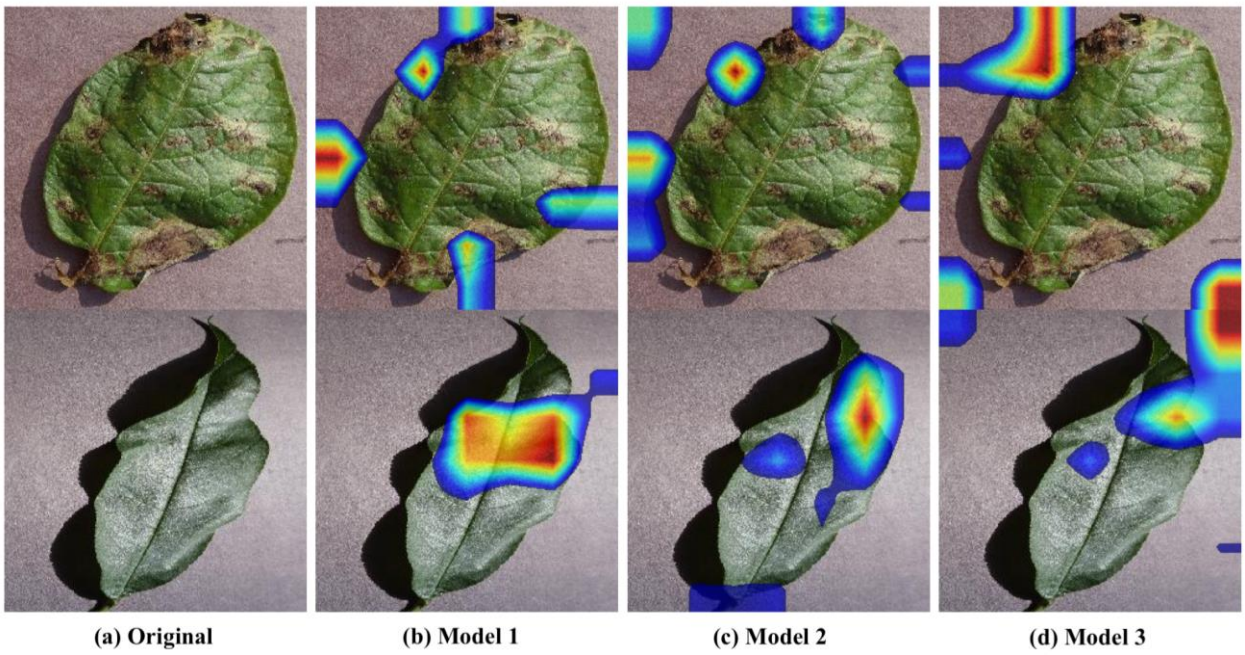

Fig. 4. Comparative Interpretability Analysis of Correctly Classified Samples.

C. Experiments 2 & 3: The Impact of Strong Regularization

In response to the identified shortcut learning, we trained a regularized version of our hybrid model (Model 2) and a pure CNN (Model 3) using a strong augmentation pipeline. Both models maintained high performance (Table I) and, demonstrated a significant improvement in diagnostic capability on the most difficult classes (Table II and Fig. 6).

However, the interpretability analysis of the regularized models reveals that this improved performance does not represent a complete solution. As shown in the heatmaps (Fig. 4), activations often diffuse across the leaf surface rather than pinpointing specific lesions, and in some cases, still activate on background elements. This persistent behavior leads to our primary conclusion: while strong regularization demonstrably improves a model's focus and diagnostic capability on hard cases, the shortcut learning problem is deeply embedded within the dataset itself and is not fully resolved by this training strategy alone.

V. Discussion

The central finding of this research is not the development of a model with incrementally higher accuracy, but the demonstration that high performance on a widely-used benchmark like PlantVillage can be fundamentally decoupled from genuine diagnostic reasoning. Our systematic, investigation provides evidence that state-of-the-art deep learning models, irrespective of their architecture, are highly susceptible to exploiting spurious correlations.

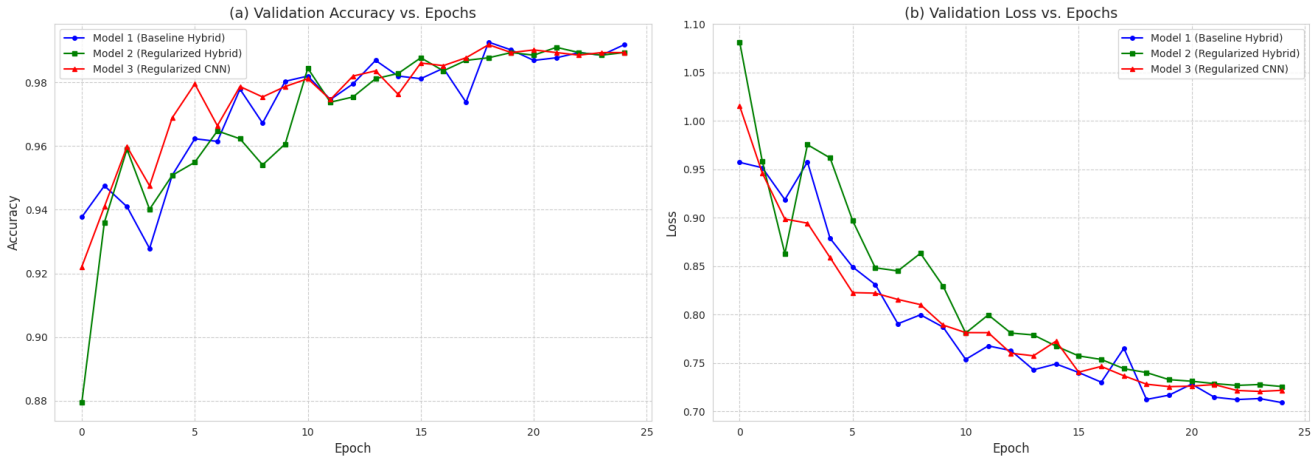

Fig. 5. Comparative Validation Performance

.

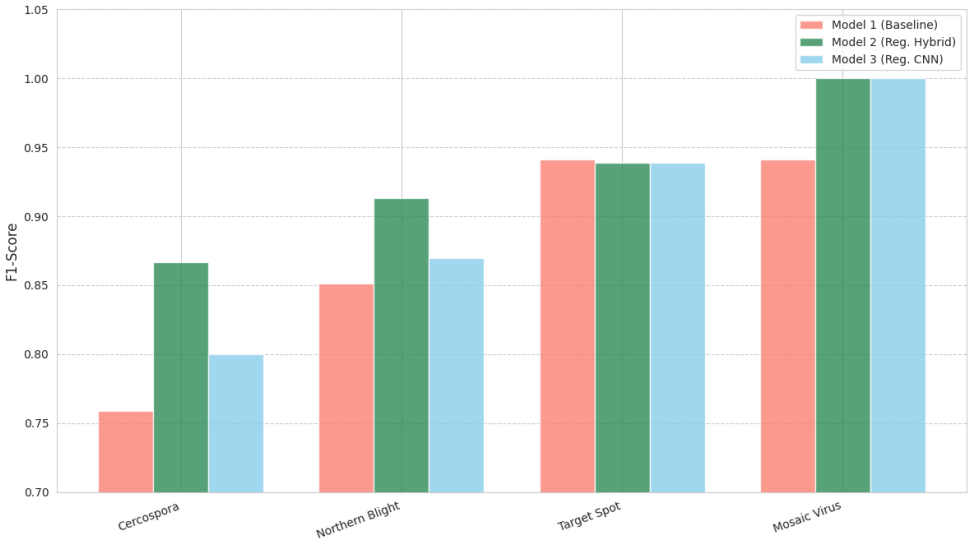

Fig. 6. F1-Score Comparison on Diagnostically Challenging Classes.

The initial 98.61% accuracy of our baseline model, while seemingly impressive, serves as a clear example of this discordance. The interpretability analysis revealed that the model learned an effective but scientifically invalid classification strategy. The fact that this behavior persisted, albeit to a lesser degree, even after applying strong regularization with RandomErasing, strongly suggests that the issue is deeply embedded within the dataset itself. These shortcuts, likely stemming from consistent imaging conditions (e.g., lighting, background surfaces) for certain classes, provide "easier" statistical patterns for the model to learn than the complex and variable patterns of actual diseases. As theorized by Geirhos et al. [6], our models were not failing; they were rationally optimizing for the features that provided the most reward.

This finding places our work in direct conversation with critical investigations in other high-stakes domains. The shortcut behavior we observed in the PlantVillage dataset is conceptually identical to the reliance on imaging artifacts reported in medical datasets for skin cancer and brain tumor diagnosis [7], [8]. This parallel underscores a systemic issue in the development of benchmark datasets: the pursuit of clean, controlled data can inadvertently introduce biases that compromise the real-world applicability of the models trained on them. As argued by Scheuerman et al.[17], the values embedded in the dataset creation process, such as prioritizing efficiency over contextual richness, can lead to such unintended consequences.

The role of the regularization strategy should be interpreted carefully. While RandomErasing did not completely eliminate shortcut learning, it proved to be a significant step in the right direction. By forcing the model to disregard random parts of the image, it successfully encouraged a shift in focus towards more robust features. This is evidenced by the substantial performance gains on the most diagnostically challenging classes (Table II and Fig. 6). This finding is perhaps the most pragmatically important result of our study: even if shortcuts cannot be entirely removed, robust training methods can force a model to become a "better" diagnostician where it matters most, on the ambiguous cases. Thus, RandomErasing should be seen not as a panacea, but as an essential tool for mitigating the risks of dataset bias.

TABLE II

DETAILED COMPARISON OF F1-SCORES ON THE MOST CHALLENGING CLASSES (F1-SCORE < 0.95 IN BASELINE MODEL). THE DATA CONFIRMS THE SUPERIOR DIAGNOSTIC ABILITY OF THE REGULARIZED MODEL IN DISTINGUISHING BETWEEN VISUALLY SIMILAR DISEASES..(FOR BREVITY, CLASS NAMES ARE ABBREVIATED (E.G., 'CORN: CERCOSPORA...' REFERS TO 'CORN (MAIZE) CERCOSPORA LEAF SPOT GRAY LEAF SPOT').)

Challenging Class	**Model 1 (Baseline) F1-Score**	**Model 2 (Regularized Hybrid) F1-Score**	**Model 3 (Pure CNN) F1-Score**	**Performance Change (Model 2 vs. 1)**
Corn...Cercospora...	75.86%	**86.67%**	80.00%	**+10.81%**
Corn...Northern Leaf Blight	85.11%	**91.30%**	86.96%	**+6.19%**
Tomato___Target_Spot	**94.12%**	93.88%	**93.88%**	-0.24%
Tomato___Tomato_mosaic_virus	94.12%	**100.00%**	**100.00%**	**+5.88%**

These findings have significant implications for the deployment of AI in real-world agricultural settings. A model that has learned to rely on background cues on a controlled dataset is likely to fail catastrophically when faced with the uncontrolled visual environments of a real farm. This elevates the role of XAI from a mere visualization technique to a critical validation step. Our work supports the argument made by Lapuschkin et al. [10] and others that interpretability analysis should be an indispensable part of the model development lifecycle to unmask such flawed reasoning before deployment.

Finally, the comparison between the hybrid and pure CNN models revealed that architectural complexity was not the primary solution. The simpler EfficientNet-B0, when properly regularized, performed comparably to our more complex hybrid. This underscores our main argument: the path to trustworthy AI in this domain may lie less in designing ever-more-complex architectures and more in developing (1) robust training strategies that mitigate the effects of bias, and (2) more challenging, bias-aware benchmark datasets that better reflect real-world conditions.

VI. CONCLUSION

This study began with the goal of creating a trustworthy plant disease diagnostic tool and concluded with a critical insight into the roles of architecture, training strategy, and dataset bias. We demonstrated that multiple deep learning architectures can achieve over 98% accuracy on a standard benchmark, but that this accuracy is a dangerously misleading metric of true capability. Our primary contribution is the XAI-driven evidence that these models are prone to shortcut learning, relying on spurious background cues rather than valid pathological symptoms.

While our proposed strong regularization strategy did not offer a complete cure, it significantly improved the model's diagnostic performance on the most difficult cases and steered its focus toward more relevant features. This confirms that robust training is a key component in mitigating the risks of dataset bias. We conclude that for high-stakes domains like agriculture, accuracy scores reported without accompanying interpretability validation should be considered incomplete and potentially unreliable. The future of this field must prioritize a paradigm where model trustworthiness is valued as highly as its performance. This requires a collective effort to both develop more challenging, realistic datasets and to establish rigorous, XAI-centered evaluation protocols as the standard for validating AI-driven diagnostic tools.

REFERENCES

[1] L. Li, S. Zhang, and B. Wang, "Plant disease detection and classification by deep learning—a review", IEEE Access, Vol. 9, pp. 56683-56698, 2021.

[2] P. Deepalakshmi and K. Lavanya, "Plant leaf disease detection using cnn algorithm", Int. J. Inf. Syst. Model. Des., Vol. 12, No. 1, pp. 1-21, 2021.

[3] J. Lu, L. Tan, and H. Jiang, "Review on convolutional neural network (cnn) applied to plant leaf disease classification", Agriculture, Vol. 11, No. 8, pp. 707, 2021.

[4] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis", Comput. Electron. Agric., Vol. 145, pp. 311-318, 2018.

[5] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection", Frontiers in plant science, Vol. 7, pp. 215232, 2016.

[6] R. Geirhos *et al.*, "Shortcut learning in deep neural networks", Nature Machine Intelligence, Vol. 2, No. 11, pp. 665-673, 2020.

[7] M. Nauta, R. Walsh, A. Dubowski, and C. Seifert, "Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis", Diagnostics (Basel), Vol. 12, No. 1, 2021.

[8] D. Wallis and I. Buvat, "Clever hans effect found in a widely used brain tumour mri dataset", Medical Image Analysis, Vol. 77, pp. 102368, 2022.

[9] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: An analytical review", Wiley Interdiscip. Rev. Data Min. Knowl. Discov., Vol. 11, No. 5, pp. e1424, 2021.

[10] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn", Nature Communications, Vol. 10, No. 1, pp. 1096, 2019.

[11] P. Iyenghar, "Clever hans in the loop? A critical examination of chatgpt in a human-in-the-loop framework for machinery functional safety risk analysis", Eng, Vol. 6, No. 2, pp. 31, 2025.

[12] C. Pal, S. Karmakar, I. Mukherjee, and P. P. Chakrabarti, "A lightweight and explainable cnn model for empowering plant disease diagnosis", Sci. Rep., Vol. 15, No. 1, pp. 30720, 2025.

[13] M. Arora, A. Santra, D. Pathak, M. Agrawal, S. Chopra, and H. Vachhani, "Interpretable deep learning for sustainable agriculture: Cnn and lime-based plant disease diagnosis", International Journal of Environmental Sciences, Vol. 11, No. 9s, pp. 1016-1030, 2025.

[14] W. Shafik, A. Tufail, L. C. De Silva, R. A. A. Haji Mohd Apong, and K.-H. Kim, "Deep learning technique for plant disease classification and pest detection and model explainability elevating agricultural sustainability", BMC Plant Biol., Vol. 25, No. 1, pp. 1491, 2025.

[15] B. Kaler and A. Kaur, "A systematic survey on explainable artificial intelligence (xai) for plant health monitoring: Challenges and opportunities", Applied Intelligence, Vol. 55, No. 12, pp. 889, 2025.

[16] S. Fabbrizzi, S. Papadopoulos, E. Ntoutsi, and I. Kompatsiaris, "A survey on bias in visual datasets", Computer Vision and Image Understanding, Vol. 223, pp. 103552, 2022.

[17] M. K. Scheuerman, A. Hanna, and R. Denton, "Do datasets have politics? Disciplinary values in computer vision dataset development", Proc. ACM Hum.-Comput. Interact., Vol. 5, No. CSCW2, pp. Article 317, 2021.

[18] D. Hughes and M. Salathé, "An open access repository of images on plant health to enable the development of mobile disease diagnostics", arXiv preprint arXiv:1511.08060, 2015.

[19] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation", Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 07, pp. 13001-13008, 2020.

[20] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," presented at the Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research, 2019. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html.

[21] S. Mehta and M. Rastegari, "Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer," in International Conference on Learning Representations, 2021.

مجله علمی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد *، شماره *، فصل **۱۴، صفحه **-**
دریافت: **/**/**، بازنگری: **/**/**، پذیرش: **/**/**

به سوی تشخیص قابل اعتماد: چارچوبی تفسیرپذیر برای آشکارسازی یادگیری میان‌بُر در مدل‌های تشخیص بیماری گیاهی

بابک مسعودی* [۱]، صفورا بزی [۲]
*۱- گروه مهندسی فناوری اطلاعات، دانشگاه پیام نور، تهران، ایران.
۲- گروه زیست‌شناسی، دانشگاه پیام نور، تهران، ایران.
[1*]B.masoudi@pnu.ac.ir, [2]S.bazzi@pnu.ac.ir

* نشانی نویسنده مسئول: بابک مسعودی، تهران، بلوار ارتش، دانشگاه پیام نور، گروه مهندسی فناوری اطلاعات، صندوق پستی: ۴۶۹۷-۱۹۳۹۵.

چکیده- پیاده‌سازی هوش مصنوعی در کشاورزی، نیازمند مدل‌هایی است که نه تنها دقیق، بلکه شفاف و قابل اعتماد باشند. در حالی که مدل‌های یادگیری عمیق در مجموعه داده‌های استاندارد به عملکرد چشمگیری دست یافته‌اند، اعتبار داخلی و قابلیت تعمیم‌پذیری آنها اغلب به‌جای راستی‌آزمایی، مفروض انگاشته می‌شود. این پژوهش، یک تحقیق روش‌شناختی در مورد شکاف میان دقت بالا و قابلیت اطمینان تشخیصی واقعی، به ویژه در بستر مجموعه داده‌های کنترل‌شده و آزمایشگاهی، ارائه می‌دهد. تحقیق ما با توسعه یک معماری هیبریدی آغاز شد که به دقت بالای ٪۹۸,۶۱ در مجموعه داده PlantVillage دست یافت. با این حال، یک تحلیل تفسیرپذیری متعاقب با استفاده از Layer-GradCAM نشان داد که موفقیت مدل به طور قابل توجهی بر ویژگی‌های جعلی پس‌زمینه استوار است؛ رفتاری که عموماً با عنوان «یادگیری میان‌بُر» شناخته می‌شود. برای بررسی نظام‌مند این پدیده، ما آزمایش‌های تکمیلی با مدل‌های قویاً قاعده‌مندسازی‌شده، شامل یک مدل پایه CNN، انجام دادیم. اگرچه این مدل‌ها دقت بالایی را حفظ کردند، پدیده یادگیری میان‌بُر، هرچند کاهش یافته بود، اما به طور کامل حذف نشد. یافته‌های ما نشان می‌دهد که دقت بالا در مجموعه داده‌های استاندارد مانند PlantVillage می‌تواند معیاری فریبنده برای ارزیابی اعتمادپذیری در کاربردهای دنیای واقعی باشد. این پژوهش، چالش اصلی در این حوزه را بازتعریف کرده و استدلال می‌کند که اعتبارسنجی دقیق و مبتنی بر XAI باید بخشی جدایی‌ناپذیر از فرآیند ارزیابی مدل باشد و بر تلاش برای افزایش جزئی دقت در مجموعه داده‌های اریب، اولویت یابد. در نهایت، این مقاله یک تحلیل انتقادی از محدودیت‌های چنین مجموعه داده‌هایی ارائه داده و بر نقش اساسی تفسیرپذیری در توسعه هوش مصنوعی قابل اتکا برای کاربردهای کشاورزی تأکید می‌ورزد.

واژه‌های کلیدی: هوش مصنوعی قابل تفسیر (XAI) ، یادگیری میان‌بُر، هوش مصنوعی قابل اعتماد، یادگیری عمیق، علوم گیاهی محاسباتی، تفسیرپذیری مدل.